PRODUCTION NOTE

LIBRARY TRENDS
VOLUME 52, NO. 2, 203–371
FALL 2003

Organizing the Internet

Andrew G. Torok

Issue Editor

UNIVERSITY OF ILLINOIS
GRADUATE SCHOOL OF
LIBRARY AND INFORMATION SCIENCE

LIBRARY TRENDS

Library Trends, a quarterly thematic journal, focuses on current trends in all areas of library practice. Each issue addresses a single theme in depth, exploring topics of interest primarily to practicing librarians and information scientists and secondarily to educators and students.

Library Trends is published four times annually—in summer, fall, winter, and spring—by the Graduate School of Library and Information Science at the University of Illinois, Urbana-Champaign, 501 E. Daniel Street, Champaign, IL 61820-6211.

Subscriptions: Institutional rate is $100 per volume (plus $7 for overseas subscribers). Subscriptions for an individual are $70 (plus $7 for overseas subscribers). Registered students may subscribe for $30 (plus $7 for overseas subscribers). Individual issues are $28 (shipping included); back issues other than those from the present year are $12 (plus shipping). Claims for missing numbers should be made within six months following the date of publication. All foreign subscriptions and orders must be accompanied by payment.

Address orders to: University of Illinois Press, Journals Department, 1325 S. Oak Street, Champaign, IL 61820. For out-of-print issues, contact ProQuest Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 48106-1346. **Postmaster:** Send change of address to University of Illinois Press, 1325 S. Oak Street, Champaign, IL 61820-6903.

This journal is abstracted or indexed in *Library and Information Science Abstracts, Current Contents, Current Index to Journals in Education, Information Science Abstracts, Library Literature, PAIS,* and *Social Sciences Citation Index.*

Procedures for Proposing and Guest Editing an Issue of *Library Trends*

We encourage our readers to submit ideas for future *Library Trends* themes; issue topics are developed through recommendations from members of the Publications Committee and from reader suggestions. We also encourage readers to volunteer to be issue editors or to suggest others who may be willing to be issue editors.

The style and tone of the journal is formal rather than journalistic or popular. *Library Trends* reviews the literature, summarizes current practice and thinking, and evaluates new directions in library practice. Papers must represent original work. Extensive updates of previously published papers are acceptable, but revisions or adaptations of published work are not sought. Although *Library Trends* is not formally peer-reviewed, guest editors invite articles for submission which are then critically reviewed by both the guest editor and journal editor. **Unsolicited articles are not accepted.**

An issue editor proposes the theme and scope of a new issue, draws up a list of prospective authors and article topics, and provides short annotations of each article's scope or else gives a statement of philosophy guiding the issue's development. Please send your ideas, inquiries, or prospectus to F. W. Lancaster, Editor, GSLIS Publications Office, 501 E. Daniel Street, Champaign, IL 61820-6211.

LIBRARY TRENDS

Fall 2003 52 (2) 203-371

Organizing the Internet

Andrew G. Torok

Issue Editor

UNIVERSITY OF ILLINOIS
GRADUATE SCHOOL OF
LIBRARY AND INFORMATION SCIENCE

Organizing the Internet

Contents

Introduction

ANDREW G. TOROK

THE THEME OF "ORGANIZING THE INTERNET" brings to mind the late 1950s folk-rock singer Jimmie Rodgers's song titled "The World I Used to Know." A great many developments have transpired in the world of information science since the seminal works of S. C. Bradford, Claude Shannon, Vannevar Bush, and numerous other pioneers. To those of us who have been in the information science field for several decades, the peek-a-boo devices such as Termatrex, Mortimer Taube's Uniterm cards, and discussion of pre- and postcoordinate indexing have given way to the world of browsers, HTML, XML, and numerous other ways of coding text and multimedia. The Internet and the World Wide Web have had a profound impact on how we go about storing and retrieving information. Document integrity has become transient, with little assurance that the location, existence, or even the content of a publication will be the same tomorrow as even a few minutes ago. We are often hard-pressed to determine if the failure to retrieve a publication is one associated with network infrastructure or the publisher. The dream of universal bibliographic control seems quite remote. By being able to bypass traditional publication channels, anyone can publish virtually at will. The situation becomes more chaotic when we consider the increasing redundancy of knowledge and the rampant proliferation of misinformation and disinformation, to say nothing of social concerns with pornography, copyright violations, and other flagrant obtrusions into personal rights. Nevertheless, it behooves the information worker and the information user to make some sense of order if good information is to remain the basis of learning and decision making, and if documents are to continue as an archive of human knowledge.

Andrew G. Torok, Professor, Department of Educational Technology, Research, and Assessment, Northern Illinois University, DeKalb, IL 60115
LIBRARY TRENDS, Vol. 52, No. 2, Fall 2003, pp. 203–208

As I reflected on writing this introduction, I began to ask myself just how far have we come from the world I used to know. The biggest paradigm change has not been that of technological development. Rather, the Internet has enabled virtually anyone with access to a computer to become intimately involved with the entire information cycle, namely, publishing, acquiring, organizing, and retrieving information, thereby bypassing information intermediaries such as indexers, reference librarians, and publishers. There is no question that the technology is vastly different from the early days of information retrieval. At the same time, the paperless office never materialized, nor are libraries being phased out as a result of the public's ability to access information directly from the desktop. More importantly, we still do not understand what constitutes information or how people make relevance judgments. Information retrieval (IR) to most searchers consists of character string matching between a query posed to a data source. In some ways, IR has even regressed, since now the trained search intermediary is no longer needed. The Internet consists of a vast unchecked sea and searching is referred to as "surfing." The issue is further complicated by the proliferation of document formats, incompatibility between generations of hardware, and questionable scalability of software. Even in doctoral seminars that I teach, I find the need to explain Boolean logic and patiently teach students how to develop search strategies, formulate queries, and even how to compute the precision of searches. While the Internet has empowered the general public to perform tasks once done by professionals, it has also created a large body of knowledge needing organization. Vocabulary control is extremely limited at best. The average Web searcher has little understanding of the search process much less a fundamental ability to determine the effectiveness or exhaustivity of a search. People rely on a limited set of search tools, especially general search engines such as Google, not realizing that less than 20 percent of all indexable documents are being accessed. Beyond that, there are many electronic text and multimedia publications that are not indexed at all by Web crawler software. This part of the Internet is called by many names, such as the Invisible Web, the Opaque Web, the Hidden Web, the Dark Web, and so on.

In all fairness, the Internet, especially the Web, is still in its infancy. Techniques for publishing, organizing, and accessing content are changing rapidly as a result of new technological developments, the competitive information marketplace, and the growing sophistication of searchers. As always, libraries are instrumental in promoting access to online publications, especially to those that belong to the invisible Web. Librarians are also educating users through the cooperative development known as information literacy. Developed by AECT (the Association for Educational Communications and Technology) and AASL (American Association of School Librarians) electronic information literacy standards are being

taught to children and teachers alike. The ACRL (Association of College and Research Libraries) supports similar standards for higher education. The dynamic nature of the Internet is going to require methods of organization way beyond the relatively static classification schemes that have served libraries for many years. New methods of organization must take into consideration more sophisticated techniques for content description in order to minimize such problems as retrieving pornography or to be able to detect plagiarism and copyright violations. Eventually the exponential growth of the Web will itself subside. The Internet is not free. Market regulations will eventually restrict the free ride enjoyed by Web publishers. Publication patterns will be easier to recognize as publication activity becomes more linear. The end result will be that users will be able to discriminate in terms of specifying what they want or avoiding the retrieval of unwanted items.

In terms of what "organization" means, I took a fairly broad approach. As in many natural systems, information on the Internet is self-organizing. For example, some search engines determine what is important to index or in what order items are viewed from a search based on link counts that point to a site. Other knowledge bases define themselves by document type, such as usenets, or come into existence by their uniqueness—blogs (Web Logs) come to mind. It seems that for many Web users, ease of use and access appear to dictate knowledge sources. At the same time, there are more organized efforts to identify and make Internet sources accessible. These efforts may simply be a subject sampler of links to relevant sites supporting a subject, area, field, or discipline. For example, the invisibleweb.com site provides classified links to Web-based databases that are not indexed by general search engines. Other sources, such as the Internet Public Library (http://www.ipl.org/ or http://www.libraryspot.com/), are portals that offer classified access to information on a much broader basis. The Open Directory project, also referred to as DMOZ, attempts to create a definitive catalog of the Web. The Open Directory is the most widely distributed database of Web content classified by humans. The Open Directory powers the core directory services for the Web's largest and most popular search engines and portals, including Netscape Search, AOL Search, Google, Lycos, HotBot, DirectHit, and hundreds of others.

Ad hoc classification systems are offered by directory search engines such as Yahoo, and other search engines like Google permit users to search by media type or document format, such as newspapers. Efforts are underway to improve basic document description beyond the limitations of HTML. Xtensible Markup Language (XML) and various permutations are but one example. In the library field, the Dublin Core Metadata Initiative (DCMI) is a notable example. Beyond large-scale efforts to identify and organize Internet content, many local efforts structure learning tools that provide quality information filtering of relevant Web information. They go

by names such as WebQuests, scavenger hunts, and Tracer Bullets. Perhaps someday these efforts will fuse into clear-cut methods of organization that lead to the development of information standards by which Web content can be created. At this time, all such projects can be construed as efforts to organize the Internet.

The purpose of this issue of *Library Trends* is to describe some of these efforts. Leading educators, librarians, and researchers have contributed articles that represent an integrated set of ideas but also serve to reflect the diversity embodied in the theme of "Organizing the Internet." The articles consist of general surveys designed to inform as well as in-depth investigations of specific issues and services.

It is appropriate to have the first article by John Carlo Bertot address the contributions and activities of libraries in a networked environment. Ever since ancient times, libraries have acted as organizers and caretakers of recorded knowledge. In addition to creating and maintaining major classification schemes such as Dewey, Library of Congress, and UDC (Universal Decimal Classification), libraries also pioneered the first major foray into electronic information retrieval. The Dialog system at the Lockheed facility in Palo Alto laid the groundwork for online searching and related software utilities that provide unique indexing capabilities for electronic files. Libraries have also contributed to knowledge organization through a variety of OPACs (Online Public Access Catalogs) and other public and technical services innovations. As libraries move away from these traditional systems grounded in service quality and outcomes frameworks, Professor Bertot discusses the challenges information professionals face in the networked environment.

To continue on the track developed by Bertot, the contribution from Adrienne Franco focuses on finding quality information on the Internet. She makes the point that librarians have long sought to select, organize, and evaluate information on the Internet. Her discussion includes the initial production of "webliographies" by librarians and then focuses on librarian-produced portals and portals with a high level of librarian participation.

Jerry D. Campbell examines portals from a more theoretical perspective. He discusses the Scholar's Portal project that builds on the need for a research library portal. Essentially, a scholar's portal (SP) describes efforts to create specialized subject portals for researchers, until such time as the Web becomes a digital library with seamless access to scholarly information. He builds on an earlier article by outlining the larger context within which SP falls.

As mentioned earlier, document organization is often by media type or even by domain name. A particularly good example of this is government information. Greg R. Notess provides a history of the government on the Web. He makes the point that the government is not only a major con-

tent provider on the Internet but also a source for the organization of the content. Patricia Diamond Fletcher continues the discussion of the government's involvement in organizing the Internet by providing a firsthand analysis of FirstGov.com based on a recent National Science Foundation–funded research project. FirstGov is the portal to U.S. government information and services. Her case study analyzes the reasons leading to the success of the portal.

Quite often the value of portals is to expose users to sources that they might not normally encounter in using general search engines. Even the best search engines index less than 20 percent of what is termed the indexable or "visible" Web. Many persons, even professional researchers, are not familiar with the invisible Web. Any discussion of organizing the Internet needs to address the invisible Web. The invisible Web consists of major databases and document formats that are not indexed by most general search engines. Less familiar, even to experienced searchers, are terms such as the "opaque Web" and the "Private Web." Chris Sherman and Gary Price discuss various permutations of the invisible Web. Their article should be of interest especially to end-users of the Web.

Classification of Web-based information is often determined by popularity, thus user preferences often prompt new methods of organization and access. Amanda Spink provides an overview of recent research exploring what we know about how people search the Web. Her paper reports selected findings from studies conducted from 1997 to 2002 using large-scale Web user data provided by Excite, AskJeeves, and AlltheWeb. The results of the research will have an impact on subsequent methods of organizing the Web according to use.

Any discussion of publication activity or use cannot avoid the topic of copyright. More than ever before, Web publishers are blatantly ignoring intellectual property rights, especially with respect to multimedia. This leads one to ask if organizers of Web publications are also contributing to copyright violations by inadvertently facilitating access to questionable material. Part of the problem lies in attempting to interpret current legislation regarding ownership of electronic publications. Rebecca P. Butler discusses implications for organizing the Internet from the viewpoints of both the owners/publishers and users. She analyzes several strands within the dilemma of the Internet and copyright. Web-based copyright issues are also addressed by Jane L. Hunter in the context of XML-based vocabularies developed to define usage and access rights associated with digital resources.

The next two contributions focus on specific aspects of organization, including discussion of metadata standards and issues of access based on document structure and content. Jane L. Hunter provides an overview of key metadata research issues and current projects and initiatives for improving our ability to discover, access, retrieve, and assimilate information on the Internet. Of particular interest to the end user is her review of

metadata search engine research. Kevin Crowston and Barbara H. Kwasnik continue the issue of vocabulary control in a somewhat different light. Their paper discusses the possibility of improving information access in large digital collections through the identification and use of document genre as a facet of document and query representation. They begin with a framework of the information retrieval problem with respect to genre and finish by outlining a research protocol that would provide guidance for identifying, using, and representing Web document genres.

Sometimes the larger efforts to make Internet documents available fail to fit the local needs of individuals. For example, a teacher in the classroom may have his/her own idea of appropriate resources to complement a lesson plan. Also, traditional methods of classification fail to reflect the constructivist paradigm popular in some educational environments. The belief is that, in order to engage students for maximum learning, there must be some way to not only identify relevant Web sites but also develop ways to explore them. Thus, educators and librarians like to develop customized resource lists that are then also made accessible to other Web users. Don E. Descy describes a variety of tools and techniques that essentially represent an ad hoc method of organizing Internet resources. He makes the point that teachers can construct Web learning environments containing safe sites for students. These can also act as quality information filters similar to the current awareness services as implemented in special libraries in the early days of automation.

In summary, the authors have addressed several dimensions surrounding efforts to organize the Internet. The contributions are of particular value because the content should be of interest to a wide spectrum of users, including librarians, educators, and academic researchers. Furthermore, many of the topics are treated in a fashion that ensures their relevance for a significantly longer period of time than that associated with most activities in a rapidly changing technological world.

World Libraries on the Information Superhighway: Internet-based Library Services

JOHN CARLO BERTOT

ABSTRACT

THE INTERNET IS NO LONGER a technology with which libraries experiment, dabble, or observe from afar. Rather, it is an integral part of library service that can take many forms—an extension of library collections and resources through licensed and/or digitized content, a gateway service through public access workstations, or a means through which customers can interact with the library through such services as digital reference. The advent of the Internet requires a reconceptualization of the information creation, dissemination, and consumption processes—and the role of libraries in these processes. Moreover, there is a need to examine our ability to engage in the assessment of network-based information services and resources as we move away from input/output evaluation approaches to those grounded in service quality and outcomes frameworks. Information professionals, and those relying on information professionals, face a number of challenges in the networked information resources and services environment. Meeting these challenges requires libraries to consider a variety of issues and strategies, several of which are presented in this article.

INTRODUCTION

The networked environment is complex and has multiple dimensions. This article focuses on selected issues that libraries face regarding service and resource delivery, management, organization, professional development, and assessment in the networked environment. It is an overview article and thus cannot address the full complexity of the impact of network-

John Carlo Bertot, Associate Professor, School of Information Studies, Florida State University, 244 Shores Building, Tallahassee, FL 32306
LIBRARY TRENDS, Vol. 52, No. 2, Fall 2003, pp. 209–227

based services and resources on the library as an institution and librarianship as a profession.

For the purposes of this article, the author defines the networked environment as the myriad of public, private, organizational, and other networks, systems, and applications used to provide users with access to electronic services and resources. These services and resources could be as simple as an online document viewed via a Web page or as complex as an electronic commerce/e-government interaction through which a user can purchase products and/or attain services such as renewal of a driver's license. In libraries, network-based services and resources can take many forms, including:

- Searching library holdings;
- Placing a hold or recalling library material;
- Making an interlibrary loan request;
- Licensing online databases, e-journals, and e-books for customer access;
- Digitizing library collections for online access;
- Providing organized Web pages that lead customers to library/nonlibrary content; and
- Providing real-time and asynchronous digital reference services.

Depending on the nature of the services or resources that libraries wish to provide their customers, libraries will need to invest in technology infrastructures that range in ability and expense and staff and customer training, in addition to considering a number of management and organizational issues that best enable the library to take advantage of such services and resources. Moreover, libraries will need to engage in evaluation activities that truly reflect the complexity of the networked environment in general and library network-based services and resources in particular.

Connectivity Background Data

This article is not about the digital divide. It is important, though, to provide some background data regarding library, school, and societal Internet connectivity and involvement:

- 98.7 percent of U.S. public libraries have an Internet connection, and 95.3 percent provide public access to the Internet (Bertot & McClure, 2002, p. 5);
- 50 percent of U.S. public libraries have Internet connectivity speeds of T1 (1.5 mbps) or greater (Bertot & McClure, 2002, p. 7);
- 99 percent of U.S. public schools have Internet connectivity, with 87 percent of instructional rooms having access to the Internet (National Center for Education Statistics, 2002, p. 3);

- 85 percent of U.S. public schools have broadband access to the Internet (National Center for Education Statistics, 2002, p. 4);[1]
- 95 percent of academic libraries have Internet connectivity according to the most recently available national data from 1998 (National Center for Education Statistics, 2001, p. 9).
- 54 percent of the U.S. population uses the Internet, though disparities exist by age, ethnicity, income, and education (National Telecommunications and Information Administration, 2002); and
- Recent research suggests that there are between 85,000 and 144,000 public computing sites across the United States, through which individuals might have access to the Internet (Williams, 2003).

Together, these data point to a nation that is increasingly online in the home and through a number of publicly accessible outlets such as libraries.

There are multiple dimensions to library Internet connectivity, from which a number of issues for libraries emanate. On the one hand, libraries need to pause for a moment and reflect upon a major accomplishment. In 1994, just 21 percent of U.S. public libraries were connected to the Internet (McClure, Bertot, & Zweizig, 1994). In less than ten years, public libraries have attained near 100 percent connectivity. This deserves some perspective: there are approximately 9,074 public library systems in the U.S. that have a total of 16,298 service outlets (typically branches, but also bookmobiles). This is a major accomplishment—one about which the library community should be proud.

Some additional, and final, statistics provide perspective on the implications for connectivity and network-based services and resources—this time from the Association of Research Libraries (ARL, 2002a, 2002b):[2]

- Expenditures for electronic resources account for an average of 16.3 percent of ARL library materials budgets;
- Collectively, ARL libraries expend more than $132 million on electronic resources, with an additional $14.66 million spent on their behalf for electronic resources through consortia purchasing arrangements;
- Expenditures for electronic serials have increased by nearly 900 percent since the 1994–95 reporting year; and
- Reference transactions have declined substantially since 1997 (down from 158,294 in 1997 to 105,087 in 2001), and circulation (of print material) is on the decline as well, down from 508,633 in 1999 to 459,335 in 2001.

One final data point may be of interest. The author conducted interviews with several database vendors and aggregators that provide services to both academic and public libraries during June 2003. These interviews

sought to determine the extent to which academic and public libraries subscribe to licensed resources. In particular, the interviews asked the aggregate expenditures (for licensed resources) for the top twenty-five individual academic and public libraries (exclusive of consortia and statewide licensing agreements). The findings: public libraries spend as much, if not more, on licensed resources as do academic libraries.

To be sure, network-based services and resources are an integral—and substantial—portion of ARL libraries. Though there may be a number of factors that contribute to the decline in use of traditional library services, it is likely the case that user access to networked information resources and services—library and nonlibrary (e.g., Google)—are having an impact on print material circulation and reference services. While difficult to extrapolate to other library types, one would expect similar data and trends.

The networked environment provides the opportunity to develop new services and resources, and to provide access to those services on a global scale. For example, libraries can digitize special, rare, or unique collections; collaborate with museums, archives, and historical societies to create unique digital content; engage in collaborative digital reference services; create electronic libraries; and expand collections without the need for additional physical space—and make these services available to the world and not just those individuals who walk into the building(s) housing such collections.

By marrying the connectivity, collections, and expenditure data with the service potential aspects of the networked environment, some substantive issues emerge. Library networked information service and resource provision require 1. assessment techniques that evaluate specifically library networked resources and services rather than approaches that combine traditional and network-based services and resources into a single form of assessment; 2. significant capital investments in technology, networking infrastructure, and continual operational costs for licensing/purchasing network-based content, services, and resources; 3. continual learning strategies and programs for library staff and users; and 4. new library management structures that include collections development, reference services, resource sharing, and other library activities.

Assessing Library Network-Based Services and Resources

In 1999, Lakos (1999) used the phrase "culture of assessment" in his discussion of the need for libraries to develop and sustain coherent and pervasive evaluation strategies regarding library service and resource provision. Briefly, Lakos argued that libraries need to create an organizational culture in which assessment is a key component to understanding the meeting space of users and libraries. This type of culture is one in which library services are under an ongoing evaluation system so as to foster con-

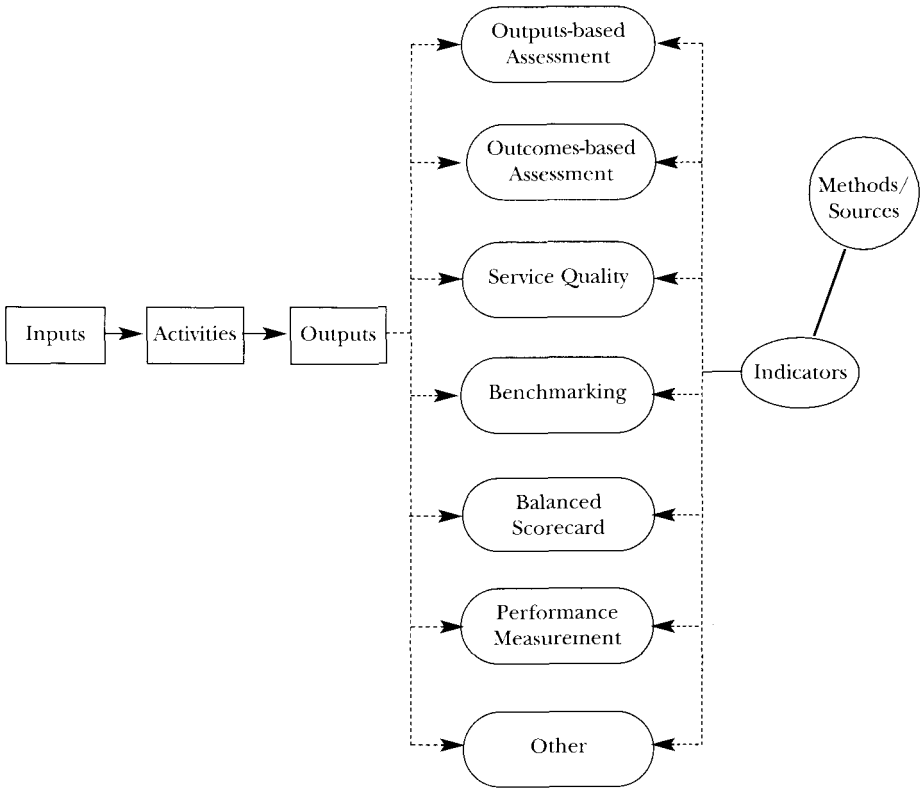
tinued improvement in meeting both library and customer needs. As will be discussed in ensuing sections of this article, such a culture requires different librarian attitudes and perceptions of library services and resources provision, different library management and working group structures, continual librarian training and education in a number of areas, and a different type of librarian than what library schools produced through their M.L.S. programs in the past.

The 1980s formalized the notion of input/output assessment techniques in librarianship (Van House et al., 1987; Van House, Weil, & McClure, 1990). This approach continues today in the networked environment as well (Bertot, McClure, & Davis, 2002; Shim et al., 2001; Bertot, McClure, & Ryan, 2000) and is in the process of incorporation of various national and international standards reviews (see, for example, the National Information Standards Organization's Z39.7 *Library Statistics* standards document at http://www.niso.org/emetrics). Indeed, entire library data collection systems center on this approach to library use, uses, and performance. For example, the Federal State Cooperative System (FSCS) managed by the National Center for Education Statistics (NCES) collects annual public library data focused on approximately fifty data elements; NCES also manages data collection activities for academic and school libraries through its library statistics program; ARL collects annual statistics from its members and so too does the Association of College Libraries (ACRL); and, as a final example, the Public Library Association collects annual statistics from a sample of public libraries through its Public Library Data Service (PLDS) program.

More recently, however, there is a push to move libraries towards service quality and outcomes assessment techniques (Hernon & Dugan, 2002; Cook & Heath, 2001). Service quality and outcomes assessment approaches differ substantially from input/output assessment but are nonetheless dependent on library inputs/outputs. Briefly (see Figure 1):

- Inputs are the resources that libraries invest (e.g., money, staff, workstations, online commercial databases);
- Activities are the library services/resources that the inputs actually generate (e.g., licensed resources availability, story hours, training sessions);
- Outputs are the service/resource results of library investments (e.g., number of users of the workstations, number of database content downloads, circulation of material);
- Outputs assessment involves the identification of the number of library activities that patrons use (e.g., number of database sessions, number of database items examined, number of training sessions conducted, etc.);
- Quality assessment involves determining the degree to which users find the library services/resources (outputs) to be satisfactory; and

Figure 1. Library Services and Assessment Frameworks.



- Outcomes assessment seeks to determine the impact of the library's services/resources (again, outputs) on the library service and resource users; or benefits, changes in skill/knowledge that library users derive from library services/resources.

Libraries that desire a comprehensive user-based assessment picture of library services/resources, therefore, need to use several evaluation strategies simultaneously—all of which are based on measures of outputs. Libraries often base their assessment strategies on trying to discover the reasons for service use/lack of use. Libraries need to know what investments (inputs) produce what services (outputs) in order to determine the perceived quality (quality assessment) and impacts (outcomes) of those services/resources. Depending on the assessed outcome and quality, library managers will want to modify their resource investment to attempt to achieve, or sustain, the desired service outcome(s). Finally, while this article focuses on issues in outcomes and service quality assessment, there

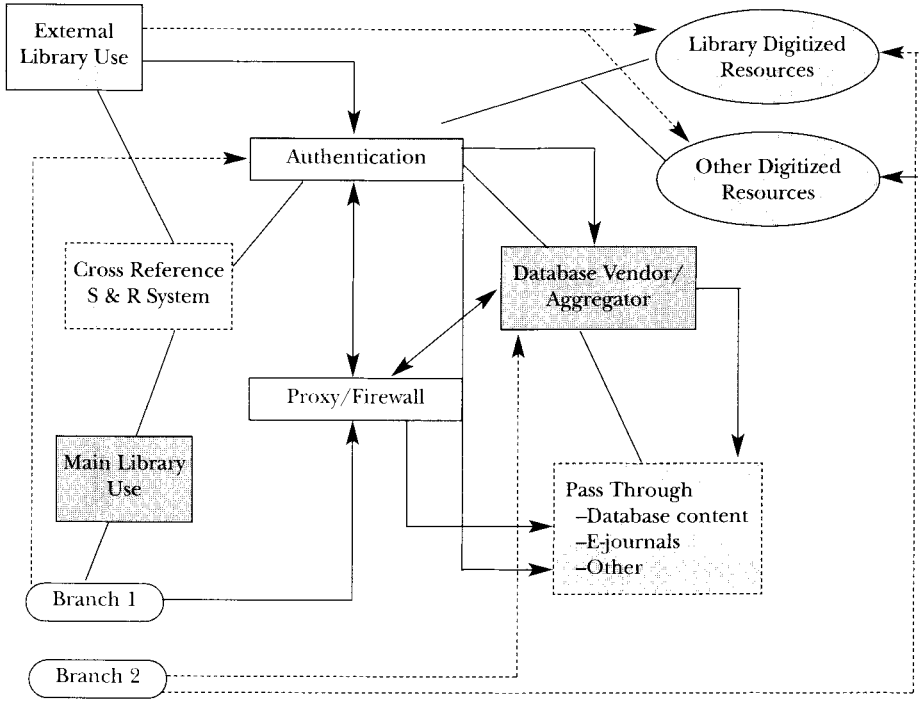
are other approaches to library services/resource evaluation that may be more appropriate (e.g., balanced scorecard) given the library's data needs and situational factors.

There are several issues associated with service quality and outcomes assessment in general and in the networked environment in particular. A more detailed discussion of these issues is available in Bertot & McClure (2003). This article, however, focuses on a high-level discussion of service quality and outcomes assessment in the networked environment. Figure 2 demonstrates the complexity of library network–based service and resource provision. At their core, service quality and outcomes assessments focus on user-based perceptions of a. the quality of library services/ resources, and b. the impacts of those services/resources on users. However, as Figure 2 shows, a vast majority of network-based services/resources that libraries provide are not under the control of the library. For example:

- Libraries are often not the content creators/managers for network-based services and resources;
  - OPACS and other internal operational software are most often purchased/leased from specific vendors and are proprietary;
  - Licensed content (e.g., databases, e-books, and the interfaces used to access vendor content) are the property of the vendor(s), and libraries typically lease that content through annual licensing agreements (though libraries can in fact purchase permanent access to e-book holdings and other resources);
  - A new, and likely to increase in use, vendor-based product is that of a cross-resource search and retrieval interface (think Google) that libraries can purchase for the purpose of enabling customers to search across vendor, Web, and library online resources through a single interface. This interface, which sits in-between the user and various other resources, is a proprietary vendor product not under the control of libraries; and
- Various technology infrastructures are not part of the library network/equipment. Customers can access "library content" from a number of locations (e.g., office, home, dorm room, other) with a wide range of computing technology and connectivity (including wireless connectivity and mobile devices). Moreover, external library connectivity has many parties involved from leased-line providers (e.g., academic computing, county information technology services, bell operating companies) to ISPs, phone lines, and wireless technologies.

To summarize, then, libraries do not control a vast majority of their network-based services and resources. Therefore, any service quality and outcomes assessment techniques will need to take that into account and ensure the account assessment of library services and resources.

Figure 2. Complexity Access to Network-Based Services and Resources.*



*An earlier version of this diagram first appeared in Bertot, McClure, & Davis (2001).

This is a particularly problematic issue with currently promoted service quality and outcomes assessment products. For example, ARL's LibQUAL+ initiative (Cook, Heath, & Thompson, 2002) and the outcomes assessment approach promoted by Hernon and Dugan (2002) use or recommend the use of survey instruments and other data collection techniques that mix online and print/traditional library services and assume library ownership of collections, services, and resources. These approaches can be quite useful at gauging library service quality/outcomes in the aggregate. Research indicates, however, that the print/traditional and electronic environments differ dramatically in important key areas such as user information-seeking behavior (Cool & Spink, 2002; Ke et al., 2002) and the ability of users to engage and extract content (Brophy, Fischer, & Clark, 2002). Lumping together traditional and networked services, therefore, leads to confounded variables, data, and results—and potentially erroneous conclusions regarding customer perceptions of outcomes and service quality.

There is a substantial need for service quality and outcomes assessment tools to probe deeper into the specifics of the services/resources they are assessing rather than continue to ask generalized questions. The general questions are helpful to provide libraries, at a glance, successful and less than successful areas of library services according to users. They do not, however, provide specific reasons for the success or lack of success of such services. Thus, libraries need to consider what the subsequent evaluation effort will be to enable in-depth probing into particular service/resource areas.

Moreover, it is likely the case that customers may actually provide feedback regarding a "library service" that is not actually provided by the library, such as online leased content. In most instances, libraries simply serve as gateways to content that resides with, and is owned by, external entities. This begs the question: Upon what, exactly, would libraries measure service quality and outcomes? For example, when a user provides feedback regarding the level of satisfaction with an online journal, is that user assessing the connectivity that leads to the journal? the interface that leads to the journal? the authentication system for access to the journal? the search interface for journal content? the journal content's format (e.g., HTML, PDF), etc.? Almost none of the above are actual services/resources provided by the library. Rather, they are particular to the various vendor systems to which the library subscribes. Asking users what they "think about a library service," therefore, is quite complex in the networked environment and points to a number of methodological problems that require resolution. Simply put, the outcomes and service quality evaluation tools of today are not adequate to engage in meaningful assessment activities for library network-based services and resources. There is much research required in this area.

Customers May Be Right, But Won't Always Get Their Way

Hernon (2002) criticizes non–user based measures of library services (e.g., input/output type measures) and strongly promotes a customer satisfaction approach to measuring the success of library services. Such a framework, adopted also by the LibQUAL+ approach, suggests that customer feedback will make its way into the resource allocation, decision-making processes, and planning activities of a library. There are two issues that emerge from this:

1. Some library services will not go away or be modified substantially regardless of user ratings. For example, the Federal Depository Library Program (FDLP) was created, among other reasons, to promote democracy and bring government closer to the people through more local dissemination and access points to government information. In the

creation of the FDLP, Congress did not specify a usage quota or user satisfaction level for such collections. This does not mean that FDLPs (or other public good–type collections such as archives and records agencies) could not benefit from user-based input. The context for such evaluation efforts, however, is important and can influence the interpretations of the results from such studies.

   a. Tangential to this issue is the notion that the Web would render the print-based FDLP program obsolete. In effect, some consider federal agency Web sites as a form of FDLP. However, since September 11th, increasing amounts of federal Web site content has been removed systematically because of national security interests. It may in fact be the case that the print-based FDLP collections, though perhaps less accessible and on a lower technology rung, are of increased significance in this era. As Patricia Diamond Fletcher discusses in this issue of *Library Trends*, FirstGov is a single point of access to online government information that continues to improve in its usability, searching, and retrieving capabilities. However, as good as FirstGov gets in terms of technology, its value decreases in direct proportion to the decrease in content to which it provides access.

2. Customer (end-user) input may have little specific impact on certain key network-based services and resources.[3] A number of key vendors have various online products and services—Elsevier has ScienceDirect, Ebsco has EbscoHost, Thompson/Gale has InfoTrac, etc. Each of these products has proprietary technology, enterprise systems, applications, interfaces, search capabilities, usage tracking capabilities, and more. The probability that a user satisfaction survey conducted on a campus library will affect the look, feel, and capabilities of each of these vendor products and services is likely remote.

While a customer-centered approach to library services in general and library network-based services in particular is desirable, it may not always yield the type of results one generally considers appropriate in a customer focus model.

Brief Discussion of Network Statistics

Much research has emerged since 1998 regarding library network statistics—essentially an input/output model for electronic library services and resources use and uses. This article does not review this work; however readers interested in such efforts should review Bertot, McClure, and Davis (2002), Shim et al. (2001), and Bertot, McClure, and Ryan (2000). For the latest in terms of network statistics data elements, definitions, and methodologies, readers are encouraged to review the NISO Z39.7 *Library Statistics* standard Web site at http://www.niso.org/emetrics.

What is of significance, however, is the notion of compliance. There are a number of forms that compliance can assume when considering network-based services and resources:

- Definitional. Groups, organizations, corporations, and individuals have expended a substantial amount of effort on the identification of network service/resource data elements and the definitions that accompany such elements. Researchers, vendor representatives, librarians, and others have worked collaboratively over the last several years through such entities as the International Standards Organization (ISO), NISO, the International Coalition of Library Consortia (ICOLC), and the Information Institute in the School of Information Studies at Florida State University to solicit library, vendor, and consortia compliance to key data elements regarding databases, online journals, and e-books.
- Reporting. Based on agreed-upon definitions, libraries and other entities (e.g., vendors) are asked to report the data regarding selected data elements in a uniform way through often centralized data reporting systems (discussed above). In general, the collection and reporting of data are executed through a decentralized process left in the hands of participating libraries with the understanding that all will adhere to the definitions as closely as possible. This approach provided various degrees of flexibility for libraries as no two libraries operate in exactly the same manner—particularly when it comes to electronic services.
- Methodological. Most library data collection and reporting efforts rely on accepted research methodologies such as focus groups, interviews, and surveys used with appropriate approaches such as sampling. Libraries are, however, left to create those surveys and/or focus group protocols to best fit the library environment in which the libraries reside—albeit with the accepted definition of elements as described above. The LibQUAL+ effort discussed before, however, requires libraries to use the same survey instrument and methodology across libraries. Thus, libraries that use the LibQUAL+ protocol also engage in methodological compliance.
- Technical. In order for libraries to offer and/or participate in the provision of various services/resources, they need to adopt a variety of technical standards such as the Z39.50 search and retrieval standard. Other standards exist or are under development—particularly in the area of metadata—that libraries will need to monitor so as to enable other services/resource provision based on those standards in the future.

To this multidimensional view of compliance, one now needs to add two more—*data* and *configuration*.

A new compliance effort—Project COUNTER (http://www.projectcounter.org/)—concentrates solely on the issue of vendor/

publisher online data compliance. Through Project COUNTER efforts, vendors and publishers have begun to adhere to a Code of Practice (http://www.projectcounter.org/code_practice.html) that will require participants to provide their usage data to a third party for data normalization efforts. The intent is to allow libraries to receive online resource usage data in a standardized format that allows comparability of data across vendors and publishers.

The COUNTER effort is a significant step forward regarding vendor/publisher online resource usage data. COUNTER largely adheres to the definitions as put forth in the ISO and NISO standards and concentrates its efforts on standardizing vendor/publisher data. The problem with COUNTER, however, is that it is quite conceivable that libraries will only be able to compare usage data *within* the library and not *across* libraries. Why? Just as no two libraries operate in the same way, no two libraries have configured their various systems and applications in the same way. While this permits a valuable degree of customization at the local institutional level that reflects a number of operational issues, it also impacts significantly what the vendors/publishers collect in terms of usage statistics (as discussed in the *Investments in Technology and Content* section of this article below). Thus while libraries may have faith in the quality of the data provided them by COUNTER-compliant vendors/publishers, comparing different library usage data (i.e., benchmarking) will likely remain the equivalent of comparing apples and oranges. Intra-library comparisons should not be a problem. If libraries want to engage in benchmarking and peer comparison activities, they will likely have to consider systems and application *configuration* compliance.

Investments in Technology and Content

The nature of the networked environment is one of rapid technological change that will necessitate *continual* investments in new technologies and upgrades to existing technology infrastructure. One-time capital investments for information technology in libraries are not a viable strategy. Libraries that wish to provide high-quality network-based services and resources to their service communities will need to develop a rational strategy and budget for the purchase, installation, maintenance, and replacement of information technology. Libraries are only beginning to recognize adequately the ongoing nature of information technology costs and to develop funding strategies to support those costs.

Beyond the need to engage in continual and regular technology investments and updates, libraries also need to consider three critical factors regarding technology and network-based services and resources:

1. The types and nature of network-based services and resources desired by libraries may require that various library technologies/systems adhere

to existing and/or emerging technical standards. For example, a library may need to comply with Z39.50 search and retrieval capabilities to provide a cross-resource search and retrieval capability (e.g., OPAC, Web, and vendor databases). This may require upgrading, the purchase of a module, or even the purchase of an entirely new OPAC so that such a system might be used by the library. This is particularly important if such a cross-resource search and retrieval system is to function in a consortia or statewide network.

2. To a large extent, library network-based resources and services are limited by the technology infrastructure of the library. For example, a library Web site requires minimally a Web server, a registered domain name, some content, and an incoming connection. If, however, libraries want to digitize and make available digitized collections via their Web sites, offer interactive services such as a "MyLibrary" feature, or conduct Web-based user surveys, libraries will need a host of additional software and equipment to engage in these activities (or at least contract with external entities for such services). It is imperative that libraries understand the relationship between their technology infrastructure and the service/resource limitations and/or capabilities that such infrastructure imposes upon the library.
3. The technology and networking infrastructures of a library determine what libraries can know about the use and uses of their network-based services and resources (Bertot, McClure, & Ryan, 2000; Shim et al., 2001). The ability of libraries to assess the use of their Web sites, as well as the ability of vendors to report the uses of database (or other) content is entirely dependent upon the library's technology installation and configuration. The use of firewalls, time-out features on workstations, and a number of other locally determined features significantly affect the nature and kinds of usage reports, and the meaning of those data, that libraries can receive and/or generate.

The above indicates the need for libraries to develop an information technology infrastructure that enables the types of network-based services and resources that they wish to provide their customers and maintain and upgrade that infrastructure regularly. Moreover, libraries need to review their technology infrastructure's capabilities continually in light of new service/resource, standards, and other developments over time. The ability for libraries to provide network-based services and resources is neither inexpensive nor a one-time proposition. It is also the case that, as technologies change, this will necessitate a change in assessment techniques that describe the use and uses of technology-based services and resources.

Content Costs and Issues

If ARL libraries are any indicator of what is happening in libraries in terms of electronic materials expenditures, then libraries are in the process

of dramatically altering their collections to the point of redefining collections development and acquisitions processes. It is not unusual for materials expenditures to change over time as new media are introduced. For example, "books/CDs on tape," video cassettes, or DVDs only became expense items for libraries as the technologies developed. The same holds true for online resources such as e-books, e-journals, and databases. Thus, there are at least three key issues regarding licensed network-based resources:

1. Libraries are increasing their licensed resources. This may occur for any number of reasons—space considerations, a way to increase collection size without significant difficulty, and/or a means through which to meet distributed customer content demand through services that are accessible from many locations. Whatever the reason, libraries are increasing the number of electronic resources to which they subscribe—and that is likely coming at the expense of other types of library material.
2. Libraries do not own many of their network-based resources. The traditional model of library collections was one of ownership—libraries bought materials that were housed in their facilities for the purpose of circulation and/or browsing by users. Until collection weeding occurred, these resources were part of a permanent collection that the library maintained. The network-based collection works quite differently, with libraries leasing content in most cases rather than owning material.[4] Thus, the expansion of electronic collections in libraries may come at the expense of collection permanency.
3. Leased collections require ongoing licensing fees. This is not a new economic model for libraries for serial-type publications that are subject to annual renewable fees.[5] However, this differs substantially as an economic model from other types of print materials, such as books, that are subject to one-time purchase fees (perhaps with periodic repurchases as material gets lost or is worn). An interesting research question that requires study is to what extent are library collections becoming leased (not owned)? Moreover, how does that evolve over time? According to the ARL data presented above, nearly 20 percent of library materials budgets is for electronic resources. It is not clear what percentage is for ongoing expenditures or what the trajectory of that expense item is—though the data point to an upward trend.

The above indicate the differing nature of materials costs and the implications for such cost considerations in the networked environment.

Library Professional and User Skills

A key question facing the library profession is "What is a librarian in the networked environment?" This seemingly simple question forces a

complex answer. While librarianship as a profession has never been monolithic in nature, the networked environment creates a situation that expands the functions of a librarian substantially. Take digital reference services as an example. Digital reference adds a series of technological, organizational, management, and knowledge layers to the reference function (Lankes et al, 2002). The library professional in the networked environment, therefore, is one who is a(n):

- *Information expert,* someone who has a fundamental understanding of information retrieval, knowledge management, information organization, information architecture and presentation, and information resource location and retrieval;
- *Communicator,* someone who has the ability to foster and exist within numerous partnerships and collaborative ventures. Librarians will also need to engage in effective communications through a variety of non–face-to-face computer mediated (CMC) forms of communication as projects may span institutions and time zones—e-mail is prevalent, but increasingly project teams use various online white board/meeting programs (e.g., Microsoft's NetMeeting), online chat, and other forms of communications technologies;
- *Instructor,* someone who can instruct users and other library staff through both formal and informal training sessions on a number of network-based services and resources (e.g., computer use, Web searching, online database use), as well as aspects of information literacy;
- *Manager,* someone who can manage varied and numerous projects, envision the possibilities of the networked environment, see the "big picture" of a project, and delegate responsibility to others;
- *Technologist,* someone who is technology savvy, is aware of new and emerging technologies, is aware of the various technology standards in existence or under development, can consider the service potential of emerging technologies, and understands a library's technology infrastructure and its implications for the ability of the library to provide various services and resources and collect use and usage data regarding those services;
- *Negotiator,* someone who is able to engage in informed contract negotiations with a number of content and resource providers such as database vendors/aggregators and systems providers. Particularly key is the ability to negotiate favorable terms for access to content (e.g., simultaneous use licenses, particular databases, desired journals/e-books) and use reporting elements and features (e.g., session counts, items accessed, searches, other);[6]
- *Strategist/Planner,* someone who thinks strategically, strives toward a vision, and can develop and implement strategic planning initiatives.

Librarians also need to engage in strategic planning activities that extend beyond the library to the larger communities that they serve, such as a university, city, county, etc.; and

- *Evaluator,* someone who is willing to benchmark and assess various initiatives—both qualitatively and quantitatively—so as to ensure project objective/goal attainment, anticipated outcomes, and service quality goals. Moreover, as evaluators, librarians will need to know the various assessment techniques available to them (e.g., network statistics/outputs, outcomes assessment, service quality), the ways in which to use these techniques so as to benefit the library's understanding of their services/resources, data analysis of the evaluation project collection activities, the interpretation of the results of such assessment approaches, and ways in which to feed the results of the evaluation projects into the library's provision of services and resources and planning activities.

While some of these qualities have long existed in the library profession, many are new and evolving. The library professional of the future is, increasingly, an information expert with a myriad of technology, management, communications, and assessment capabilities.

More significantly, perhaps, is that the education process for librarians is continual and ongoing—it is not the case that, upon graduation from a degree program, the librarian is complete in his/her education. With technology changes, new assessment tools, and various other issues, libraries need to build a continuing education process for librarians to work effectively in the evolving field of librarianship. A library degree is a necessary, but no longer sufficient, qualification for a library career. Given the skills required as outlined above, it may also be the case that "librarians" in the networked environment are more appropriately trained in disciplines (e.g., instructional design, information systems, business) other than librarianship through M.L.S. degree programs for certain library functions.

Customer Instruction

It is not possible to cover all topics in this article. It is important to mention, however, that library customers also require continual training and education regarding the networked environment in general and library network-based services and resources in particular. Indeed, libraries of all types participate in educational services that fall broadly under the header of "information literacy." Bertot and McClure (2002, p. 13) found that 42 percent of public libraries offer formal Internet/computer training courses on a variety of topics (this does not include the five- or ten-minute point-of-use sessions requested by users seeking help). Academic librarians are generally considered faculty, have teaching requirements, and often offer a wide range of "information literacy" courses that span tech-

nology and information content (Ratteray, 2002; Chiste, Glover, & Westwood, 2000).

Impact on Organizational Structure

New forms of library services require new library organizational structures (Liu, 2001). Libraries may find that function-based hierarchical structures no longer work well for library service in the networked environment. Increasingly, libraries need to consider, and in some cases are moving toward, a variety of work models, such as:

- Team-based/group activities that focus on a particular project (e.g., designing a Web site, digitizing a collection, providing a comprehensive electronic library-based collection);
- Cross-functional approaches to service development and provision that reflect the reach of network-based services. This may mean more and frequent collaboration across libraries and external library partners such as historical societies, academic units, archives, museums, and records management agencies; and
- Fluid, matrix-like structures that can quickly form to work on a project, may include a number of project subteams, and then disband upon project completion.

As such, library organizations need to consider organizational structures and management methods that better reflect their changing operating environment.

Concluding Comments

Libraries have moved beyond the use of the Internet as a novel experiment into the use and provision of network-based resources and services as a substantial—and increasing—aspect of library services. The evolution from dabbling to entrenchment has a number of library institutional, organizational, management, professional, and assessment implications that this article discussed selectively. The real work has begun, and libraries are working diligently to accommodate the new reality in innovative, strategic, and visionary ways.

This article suggests, however, that we have much to learn about library involvement with and use of network-based resources. So, too, do we have much to learn regarding customer perceptions of network-based service quality and outcomes. It is important for librarians and information professionals to focus on the capabilities enabled by the networked environment rather than the complications brought forth by the complexity of network-based information resources and services. The profession's and researcher's understanding of the networked environment will evolve through experimentation and study.

NOTES

1. "Broadband," in the National Center for Education Statistics survey of public schools, includes cable modem service, T1, Fractional T1/T3, and T3/DS3 service.
2. ARL has approximately 120 academic library members. Additional information on ARL is available at http://www.arl.org.
3. This article is not a critique of vendor systems, products, or services. Any mention of specific products/services is illustrative only.
4. Some e-book vendors do allow libraries to purchase the electronic book and add those titles to their permanent collections.
5. Even in the case of serials, however, the library owns the back issues that it purchased.
6. Readers should review the network statistics and their definitions found in the NISO Z39.7 *Library Statistics* standard found at http://www.niso.org/emetrics for additional information regarding the data elements that they may want vendors to report.

REFERENCES

Association of Research Libraries (ARL). (2002a). *ARL supplementary statistics 2000–2001.* Washington, D.C.: Association of Research Libraries. Retrieved March 30, 2003, from http://www.arl.org/stats/pubpdf/sup01.pdf.

Association of Research Libraries (ARL). (2002b). *ARL statistics 2000–2001.* Washington, D.C.: Association of Research Libraries. Retrieved March 30, 2003, from http://www.arl.org/ stats/pubpdf/arlstat01.pdf.

Bertot, J. C., & McClure, C. R. (2002). *Public libraries and the Internet 2002: Internet connectivity and networked services.* Washington, D.C.: Institute of Museum and Library Services. Retrieved March 25, 2003, from http://www.ii.fsu.edu/Projects/2002pli/2002.plinternet.study.pdf.

Bertot, J. C., & McClure, C. R. (2003). Outcomes assessment in the networked environment: Research questions, issues, considerations, and moving forward. *Library Trends, 51*(4), 590–613.

Bertot, J. C., McClure, C. R., & Davis, D. (2001). *Developing a national data collection model for public library networked statistics and performance measures: Interim report.* Washington, D.C.: Institute of Museum and Library Services. Retrieved March 25, 2003, from http://www.ii.fsu.edu/Projects/IMLS/interim.report.may2001.pdf.

Bertot, J. C., McClure, C. R., & Davis, D. (2002). *Developing a national data collection model for public library networked statistics and performance measures: Final report.* Washington, D.C.: Institute of Museum and Library Services.

Bertot, J. C., McClure, C. R., & Ryan, J. (2000). *Statistics and performance measures for public library networked services.* Chicago: American Library Association.

Brophy, P., Fisher, S., & Clark, Z. (2002). *Libraries without walls 4: The delivery of library services to distant users.* London: Facet Publishing.

Chiste, K. B., Glover, A, & Westwood, G. (2000). Infiltration and entrenchment: Capturing and securing information literacy territory in academe. *Journal of Academic Librarianship, 26*(3), 202–208.

Christensen, C. M. (1997). *The innovator's dilemma.* Boston: Harvard Business School Press.

Cook, C., & Heath, F. M. (2001). Users' perceptions of library service quality. *Library Trends, 49*(4), 538–584.

Cook, C., Heath, F. M., & Thompson, B. T. (2002). Score norms for improving library service quality: A LibQUAL+™ study. *Portal: Libraries and the Academy, 2*(1), 13–26.

Cool, C., & Spink, A. (2002). Issues of context in information retrieval (IR): An introduction to the special issue. *Information Processing & Management, 38*(5), 605–611.

Hernon, P. (2002). Outcomes are key but not the whole story. *Journal of Academic Librarianship, 28*(1–2), 54–55.

Hernon, P., & Dugan, R. E. (2002). *An action plan for outcomes assessment in your library.* Chicago: American Library Association.

Ke, H. R., Kwakkelaar, R., Tai, Y. M., & Chen, L. C. (2002). Exploring behavior of e-journal users in science and technology: Transaction log analysis of Elsevier's ScienceDirect OnSite in Taiwan. *Library and Information Science Research, 24*(3), 265–291.

Lakos, A. (1999). The missing ingredient: Culture of assessment in libraries. *Performance Measurement and Metrics, 1*(1), 3–7. Retrieved March 20, 2003, from http://www.aslib.com/pmm/1999/aug/opinion.pdf.

Lankes, R. D., McClure, C. R., Gross, M., & Pomerantz, J. (2002). *Implementing digital reference services: Setting standards and making it real.* New York: Neal-Schuman Publishers.

Liu, L. G. (2001). *The role and impact of the Internet on library and information services.* Westport, CT: Greenwood Press.

McClure, C. R., Bertot, J. C., & Zweizig, D. L. (1994). *Public libraries and the Internet: Study results, policy issues, and recommendations.* Washington, D.C.: National Commission on Libraries and Information Science.

National Center for Education Statistics. (2001). *Academic libraries: 1998.* [NCES 2001–341]. Washington, D.C.: U.S. Department of Education, National Center for Education Statistics. Retrieved March 10, 2003, from http://nces.ed.gov/pubs2001/2001341.PDF.

National Center for Education Statistics. (2002). *Internet access in U.S. public schools and classrooms: 1994–2001.* Washington, D.C.: U.S. Department of Education, National Center for Education Statistics. Retrieved March 10, 2003, from http://nces.ed.gov/pubs2002/2002018.pdf.

National Telecommunications and Information Administration. (2002). *A nation online: How Americans are expanding their use of the Internet.* Washington, D.C.: Department of Commerce, National Telecommunications and Information Administration. Retrieved January 15, 2003, from http://www.ntia.doc.gov/ntiahome/dn/html/anationonline2.htm.

Ratteray, O. M. (2002). Information literacy in self-study and accreditation. *Journal of Academic Librarianship, 28*(6), 368–375.

Shim, J. W., McClure, C. R., Fraser, B. T., Bertot, J. C., Dagli, A., & Leahy, E. H. (2001). *Measures and statistics for research library networked services: Procedures and issues: ARL e-metrics phase II report.* Washington, D.C.: Association of Research Libraries. Retrieved February 15, 2003, from http://www.arl.org/stats/newmeas/emetrics/index.html.

Van House, N. A., Lynch, M. J., McClure, C. R., Zweizig, D. L., & Rodger, E. J. (1987). *Output measures for public libraries: A manual of standardized procedures.* 2nd ed. Chicago: American Library Association.

Van House, N. A., Weil, B. T., & McClure, C. R. (1990). *Measuring academic library performance: A practical approach.* Chicago: American Library Association.

Williams, K. (2003). Research note: Across the United States, 85,000 to 144,000 public computing sites. *First Monday, 8*(4). Retrieved April 8, 2003, from http://firstmonday.org/issues/issue8_4/williams/index.html.

Gateways to the Internet: Finding Quality Information on the Internet

ADRIENNE FRANCO

ABSTRACT

Librarians have long sought to select, evaluate, and organize information on the Internet. Efforts began with individual librarians sharing bookmark files of favorite sites and progressed to increasingly large, collaboratively produced general and subject/discipline-specific gateway Web sites or megasites. Megasites list major resources usually in a particular subject area or discipline. Library portals that review, evaluate, and sometimes rate and rank resources grew from some of these Web sites. Both megasites and portals serve as gateways to the Internet. Many portals have developed from relatively small static files into large, dynamically generated databases providing descriptive annotations of selected resources and are increasingly overseen as global projects with formal policies and procedures. Portals now provide increasingly complex and sophisticated browse and search capabilities with a multitude of access points, often including call numbers and subject headings. These are described and compared. Future trends such as increased collaboration among portals; automated location, selection, and cataloging of resources; integration of multiple resource types; and increased access to full-content and virtual library services are also discussed.

INTRODUCTION

Librarians have long been involved in efforts to select, organize, describe, and evaluate Internet resources. Librarian-produced Internet tools have much to offer that commercial search engines and other tools lack:

Adrienne Franco, Reference and Instructional Services Librarian, Iona College Libraries, 715 North Avenue, New Rochester, NY 10801–1890
LIBRARY TRENDS, Vol. 52, No. 2, Fall 2003, pp. 228–246

> While these search engines [Yahoo and Alta Vista] and others like them have strengths, their weaknesses are well known: a high percentage of nonauthoritative content mixed with quality content that, when indexed together, makes locating relevant information serendipitous at best. (Wells et al., 1999, p. 347)

Early on, individual librarians compiled bookmark files that listed favorite sites. These lists often reflected institutional priorities and usually had a limited geographical focus as well. In fact, the well-respected Librarian's Index to the Internet began as then Berkeley Public Library librarian Carole Leita's gopher bookmark file (Buchwald, 2002, p. 38). As the Internet grew in size and audience and became more accessible, librarians worked collaboratively to create and maintain resource sites and megasites. These might be multidisciplinary, as in selections of general reference resources, or subject or discipline specific. Initially, following print models of bibliographic control, these guides were essentially Web bibliographies or "Webliographies." Megasites (sometimes called "metasites") are larger and more comprehensive. Webliographies and megasites became increasingly sophisticated, providing descriptive annotations. Portals are larger still and often evaluate and sometimes rate megasites and other Internet resources.

The LITA Internet Portals Interest Group

> defines a portal as a service (and related systems and approaches to organization) that facilitates organized knowledge discovery via information accessible through the Internet. (American Library Association. Library and Information Technology Association, n.d.)

Portals are now often supported as independent projects and are frequently underwritten financially through state, local, or national governments or private philanthropic funding (cf., for example, Ansdell, 2000; Buchwald, 2002, p. 38; Wells et al., 1999, p. 347).

As portals became more established and grew larger, librarians took advantage of software advances to convert them into databases that are browsable and searchable by multiple access points, frequently including call numbers and subject headings.

Scope

This article will focus primarily on librarian-produced portals or portals with a high level of librarian participation. Sites described and discussed are freely available on the Web. These portals will be described and compared. Excluded or de-emphasized are sites created and maintained primarily outside the library community, print resources including books and articles, information available only in fee-based subscription databases, and search engines.

Article Background

This article grew out of a presentation given on October 14, 1999, by the author and a colleague, Richard Palladino, at the 10th Annual Meeting of the International Information Management Association (IIMA) held at Iona College. An invitation to participate in this conference was extended to Iona College faculty and staff. The concept of information management seemed especially pertinent to librarians and the opportunity to present before an audience of nonlibrarians was especially intriguing and attractive. Aware of widespread concern about the quality (or lack of quality) on the World Wide Web, thoughts of librarians extending bibliographic and quality control from print to the Web came to mind, and so we decided to share this with our fellow information professionals. The Web page "Finding Quality Information on the World Wide Web" (http://www.iona.edu/faculty/afranco/iima/webliog.htm) was created for presentation at the conference and has been maintained since then and most recently updated on April 4, 2002. We were the only librarians to present at this conference. Information professionals from around the world attended, and their feedback was overwhelmingly positive. Some took us aside and said they had been unaware of librarians' attempt to select, organize, and evaluate Internet resources.

Finding Subject Guides and Megasites

Finding the Newest Quality Sites

Although subject guides and megasites are included in the portals discussed in this article, newer resources may not yet be included. Methods that are described here are often also used by librarians at portal sites to find resources to be considered for review and inclusion.

Subject guides and megasites are often created under the auspices of organizations such as college and university academic departments, government agencies, nonprofit organizations, professional associations, trade associations, and corporations, as well as libraries. Some are the product of special, highly structured projects while others may represent the efforts of individuals or informal groups. For example, a university biology faculty member or librarian may create a Webliography of favorite sites.

Methods used to find quality sites include:

- Mailing lists and discussion groups for resource announcements and recommendations;
- Print sources such as books and journal, magazine, or newspaper articles;
- Search engines, using carefully constructed search queries. Such queries may include terms that describe a discipline or broad subject area as well as words such as "resources," "megasites," "Webliography," "Internet," etc. For example:

biology + megasites
biology + "internet resources"
biology + "information resources"
biology + webliography (or, biology + bibliography)

It may be helpful to limit searches to the titles of Web pages only and possibly to domains such as .edu, .gov, or .org to retrieve megasites produced by academic institutions, libraries, nonprofit organizations, or government agencies. One can exclude domains if desired as well, e.g. exclude ".com." Of course, you will have to screen search results yourself.

- Other strategies for locating megasites include the following:
  - Determine which academic institutions have degree programs in a particular field or discipline. (To help you identify which institutions have programs in a particular field, consult print or electronic directories, e.g., College Blue Book or Peterson's college guides);
  - Once you've identified an appropriate institution, try using the url: "www.universityname.edu" (for U.S. universities), or use a Web directory such as: American Universities (http://www.clas.ufl.edu/CLAS/american-universities.html);
  - Look for appropriate academic department page(s) as well as library page(s);
  - Look for Web documents that may include such title words/ terms as "Links," "Resources," "Web Sites," etc.

Major Web Rating and Evaluation Portal Sites

Eventually, quality megasites will be accessible through portals such as the Librarian's Index to the Internet and Infomine. Specific portals that are described and compared in this article include Librarians' Index to the Internet, Infomine, Internet Public Library, MEL (Michigan Electronic Library), BUBL Link 5:15, Internet Scout Project, and Academic Info. These are described and compared in Tables 1–7.

Comparing the data in these tables, we see commonalities but also significant differences. For example, most provide at least basic keyword search capabilities and at least minimal annotations. Most also began in the early to mid-1990s and provide selected sites, though criteria are not always explicitly stated on their Web sites.

Differences among them, however, are significant, so users are advised to not limit their searches for quality resources to a single portal. Examples of major differences include: primary audience, level of detail in records, number of access points, presence or absence of controlled vocabulary and classification system numbers, degree of searchability and browsability, and comprehensiveness of annotations.

For example, primary audiences range from public library users (Librarians' Index to the Internet) to academics (Infomine and Academic Info) and all the Internet community (Internet Public Library, MEL).

Table 1.

Name of Web Rating and Evaluation Site:	Librarians Index to the Internet
Site URL:	http://lii.org
Mission Statement, Description, Audience:	"The mission of Librarians' Index to the Internet is to provide a well-organized point of access for reliable, trustworthy, librarian-selected Internet resources, serving California, the nation, and the world."
Year Founded:	1990
Origins/History:	Began as librarian Carol Leita's gopher bookmark file
Approximate Number of Records:	Over 10,000 as of end of 2002
Selection Criteria:	Detailed criteria described at: http://lii.org/search/file/pubcriteria. Free sites or sites that offer significant free content only are included. Evaluation criteria include authority, scope and audience, content, design, function, and shelf life.
Annotations?	YES
Sites Rated? (e.g., with graphics such as stars)	NO
Browsable?	By hierarchical terms, general to specific. By LC subject headings from advanced search screen.
Searchable?	YES, with fully-functional search engine
Classification System Used?	NO
Subject Headings/Controlled Vocabulary?	LCSH
E-Mail Announcements/ Alerts for New Sites Added?	YES
Staffing:	4 part-time staff including a cataloger, 2 editors, and a computer programmer plus more than 100 volunteer indexer librarians
Responsible Person(s)/ Institution(s):	Library of California, Karen G. Schneider
Funding and Support:	Library of California, grants such as LSTA
Hosted by:	UC Berkeley SunSITE
Prime URL for "about" information:	http://lii.org/search/file/about
COMMENTS:	Although emphasis is on public libraries, resources and annotations are useful for academics as well.

Table 2.

Name of Web Rating and Evaluation Site:	INFOMINE: Scholarly Internet Resource Collections
Site URL:	http://infomine.ucr.edu/
Mission Statement, Description, Audience:	"INFOMINE is a virtual library of Internet resources relevant to faculty, students, and research staff at the university level. It contains useful Internet resources such as databases, electronic journals, electronic books, bulletin boards, mailing lists, online library card catalogs, articles, directories of researchers, and many other types of information." Scope information available at: http://infomine.ucr.edu/about/scope.php
Year Founded:	1994
Origins/History:	Begun by librarians at the University of California, Riverside. Librarians from other academic institutions now participate as well. Infomine is now a cooperative project.
Approximate Number of Records:	Over 40,000; half selected by librarian "experts"; the other half by robot crawlers (Mitchell, 2003).
Selection Criteria:	"University level research and educational tools on the Internet."
Annotations?	YES
Sites Rated? (e.g., with graphics such as stars)	Graphical symbols used to distinguish, for example, librarian-selected records.
Browsable?	From main screen by hierarchical subject-specific database (e.g., Business & Economics). From advanced search screen by LC classification numbers.
Searchable?	YES
Classification System Used?	YES (LC)
Subject Headings/Controlled Vocabulary?	LCSH
E-Mail Announcements/ Alerts for New Sites Added?	YES
Staffing:	"Librarians from The University of California, Wake Forest University, California State University, The University of Detroit—Mercy, and other universities and colleges" (cf. http://infomine.ucr.edu/about/). Other libraries invited to participate.
Responsible Person(s)/ Institution(s):	Primarily University of California, Riverside
Funding and Support:	State, federal, and other grants
Hosted by:	University of California, Riverside
Prime URL for "about" information:	http://infomine.ucr.edu/about/
COMMENTS:	Now part of **LOOK (Libraries of Organized Online Knowledge,** formerly Fiat Lux), a collaborative project of multiple portal sites.

Table 3.

Name of Web Rating and Evaluation Site:	Internet Public Library
Site URL:	http://www.ipl.org/
Mission Statement, Description, Audience:	"The first public library of and for the Internet community" (cf. http://www.ipl.org/div/about/iplfaq.html). However, audience is not "public library" users but all members of the Internet community as well as librarians. Designed on a library model, IPL provides library services and resources such as Reference and links to free online books and articles. Primary focus does not seem to be Web site evaluation
Year Founded:	1995
Origins/History:	Began in winter 1995 as a project of the School of Information and Library Studies at the University of Michigan
Approximate Number of Records:	Not found at site
Selection Criteria:	Not found at site
Annotations?	YES, but seem to appear only when browsing rather than searching. Brief and often are quoted from the site itself.
Sites Rated? (e.g., with graphics such as stars)	NO
Browsable?	Yes, by hierarchical terms general to specific. Browses do retrieve records with annotations.
Searchable?	Yes, but simple searches only. Searches do not retrieve annotated records but simply a list of links.
Classification System Used?	NO
Subject Headings/Controlled Vocabulary?	NO
E-Mail Announcements/ Alerts for New Sites Added?	Not found at site
Staffing:	Sue Davidsen, Managing Director, and two other staff members. Students at the host institution. Others invited to collaborate.
Responsible Person(s)/ Institution(s):	University of Michigan School of Information
Funding and Support:	University of Michigan School of Information. Actively seeking other funding.
Hosted by:	University of Michigan School of Information
Prime URL for "about" information:	http://www.ipl.org/div/about/
COMMENTS:	Also includes original content pathfinders and documents created for IPL. Includes records formerly in the Argus Clearinghouse which was discontinued on January 23, 2002.

Table 4.

Name of Web Rating and Evaluation Site:	MEL: Michigan Electronic Library Best of the Internet Selected by Librarians
Site URL:	Main url: http://www.michigan.gov/hal/0,1607,7-160-15481_15483—,00.html http://www.michigan.gov/hal URL for "Best of the Internet": http://mel.org/melindex.html
Mission Statement, Description, Audience:	"Michigan's virtual library will link all Michigan residents to the information they need, when they need it, where they need it, and in the format they desire."
Year Founded:	1992
Origins/History:	Began as GoMLink gopher service
Approximate Number of Records:	Over 20,000
Selection Criteria:	Sites are selected that meet the needs of Michigan's libraries and citizens. The Web site alludes to specific selection criteria followed by their selectors but does not include them. "Collection Policy for the Michigan eLibrary—Best of the Internet," http://mel.org/about/melcollection.html
Annotations?	YES, but very brief and not for all records. Some are quotes from linked sites.
Sites Rated? (e.g., with graphics such as stars)	NO
Browsable?	YES, by hierarchical terms general to specific.
Searchable?	YES, but simple search only. Seems to be keyword access only. No advanced search features (e.g., limiting).
Classification System Used?	NO
Subject Headings/Controlled Vocabulary?	NO
E-Mail Announcements/ Alerts for New Sites Added?	NO
Staffing:	11 manager/selector librarians
Responsible Person(s)/ Institution(s):	Michigan State Library
Funding and Support:	Michigan State Library, LSTA "via the Institute of Museum and Library Services (IMLS)," and other grants
Hosted by:	State of Michigan
Prime URL for "about" information:	http://mel.org/about/aboutmel.html
COMMENTS:	Best of the Internet is only a small part of MEL, which is a virtual library.

Table 5.

Name of Web Rating and Evaluation Site:	BUBL / Link 5:15, catalog of Internet resources (part of BUBL)
Site URL:	http://bubl.ac.uk/link/ddc.html (Dewey) http://bubl.ac.uk/link/(alternative subject interface)
Mission Statement, Description, Audience:	"Aimed towards the UK higher education academic and research community" and librarians; "a catalogue of selected Internet resources covering all academic subject areas and catalogued according to DDC."
Year Founded:	BUBL 5:15 began in March 1997. Original BUBL began in 1990.
Origins/History:	BUBL founded as BUlletin Board for Libraries, aimed at librarians. LINK stands for Libraries of Networked Knowledge.
Approximate Number of Records:	Over 11,000 resources
Selection Criteria:	"Academic relevance, up-to-date information and completeness" (cf. Williamson, 2000). Williamson also lists specific types of resources that are given priority, e.g., online books and book collections.
Annotations?	YES, descriptive
Sites Rated? (e.g., with graphics such as stars)	NO
Browsable?	By BUBL subject tree (hierarchical subjects, from general to specific) and by Dewey classification numbers
Searchable?	Fully cataloged with multiple access points. Simple and advanced search available. Fielded searching and sophisticated search features (e.g., Boolean, truncation, etc.) are available.
Classification System Used?	Dewey Decimal
Subject Headings/Controlled Vocabulary?	Enhanced LCSH
E-Mail Announcements/ Alerts for New Sites Added?	Update information available on "lis-link" mailing list (archive and subscription instructions available at http://www.jiscmail.ac.uk/lists/LIS-LINK.html). Update bulletins also available at http://bubl.ac.uk/news/updates/
Staffing:	2 full-time staff and 1 part-time staff member
Responsible Person(s)/ Institution(s):	Andersonian Library, Strathclyde University, 101 St. James Road, Glasgow G4 0NS, Scotland
Funding and Support:	Joint Information Systems Committee (JISC) of the Higher Education Funding Councils of England, Scotland and Wales and the Department of Education for Northern Ireland

Table 5. (continued)

Hosted by:	BUBL has own server
Prime URL for "about" information:	http://bubl.ac.uk/admin/
COMMENTS:	Can also limit search by file type, e.g., sound

Some are stand-alone portals (e.g., Librarians' Index to the Internet) while others are part of larger virtual libraries (e.g., Internet Public Library). It appears that the stand-alone portals are more likely to provide in-depth records, multiple access points, and more sophisticated search options than those that are only part of a virtual library. This is true, for example, when one compares Librarians' Index to the Internet to the Internet Public Library.

Current Trends

It is well documented that search engines cover only a small fraction of resources available on the Web (cf. Lawrence & Giles, n.d.). Portals cover even a smaller percentage of resources. Internet users are less aware of the portals discussed in this article and if they are aware may use them less frequently than search engines because they retrieve fewer records with each search. It is easy to confuse volume with quality of search results. The portals can offer quality that search engines, even those that increasingly use "intelligent" search algorithms, are less able to provide. Still, portal leadership has recognized the need to cover more resources. This has resulted in many trends and developments that are both current and developing. These current and developing trends are discussed below.

Automation and Software

Creation and development of sophisticated software has allowed portal sites to automate almost every aspect of their sites from collection development to record creation, search, and retrieval of information. For example, Infomine uses crawlers to find, evaluate, and select resources for inclusion. Half of their database consists of resources that are machine-selected. Other tasks increasingly automated include record creation, indexing, and even brief descriptive annotations. Automation has played a major role in virtually all of the trends that follow.

Growth

Portals such as Infomine and Librarians' Index to the Internet have been rapidly increasing the number of resources included. Consistent with increased diversity of Internet resources, portals now cover not only HTML

Table 6.

Name of Web Rating and Evaluation Site:	Internet Scout Project (offering access to weekly Scout Report, Scout Report Archives, and NSDL Scout Reports)
Site URL:	http://scout.wisc.edu/ http://scout.wisc.edu/report/sr/current/ (Scout Report, current issue) http://scout.wisc.edu/archives/ (Scout Report, archives) http://scout.wisc.edu/nsdl-reports/ (NSDL Scout Reports—National Science Digital Library)
Mission Statement, Description, Audience:	"To provide timely information to the education community about valuable Internet resources." Audience: "K-12 and higher education faculty, staff, and students, as well as interested members of the general public" (cf. http://scout.wisc.edu/about/).
Year Founded:	1994
Origins/History:	Subject-specific scout reports for Business & Economics, Social Sciences & Humanities, and Science & Engineering discontinued in 2001 due to lack of funding (cf. Search engines, 2001).
Approximate Number of Records:	Over 11,000
Selection Criteria:	Content, Authority, Information Maintenance, Presentation, Availability, and Cost. Detailed criteria listed at http://scout.wisc.edu/report/sr/criteria.html
Annotations?	YES, critical annotations (cf. http://scout.wisc.edu/archives/)
Sites Rated? (e.g., with graphics such as stars)	NO
Browsable?	By LCSH
Searchable?	Fully cataloged with multiple access points. Simple and advanced search available. Fielded searching and sophisticated search features (e.g., Boolean, truncation, phrase searching, etc.) are available.
Classification System Used?	Broad LC class only, e.g., Z, RG, etc. Not searchable or browsable.
Subject Headings/Controlled Vocabulary?	LCSH
E-Mail Announcements/ Alerts for New Sites Added?	YES. Can subscribe by going to http://scout.wisc.edu/report/sr/srsubscribe.html
Staffing:	17 staff including 2 librarian catalogers. Sites selected by "professional librarians, educators, and content specialists."

Table 6. (continued)

Responsible Person(s)/ Institution(s):	Department of Computer Science, University of Wisconsin-Madison
Funding and Support:	National Science Foundation
Hosted by:	University of Wisconsin
Prime URL for "about" information:	http://scout.wisc.edu/about/
COMMENTS:	Links regularly checked and updated. Scout Portal Toolkit software information available at http://scout.wisc.edu/research/SPT/.

but other file types as well, including PDF, images, and multimedia. Half of Infomine's 40,000 records are machine generated, with the other half created by librarian experts.

Static Files to Databases

As content has increased, most portals have converted from static files to databases with multiple access points and sophisticated searching capabilities to facilitate searching and retrieval of records.

Organization and Structure

Portals have developed into highly organized and structured projects increasingly supported by government and philanthropic agencies. Many have become independent organizations financed separately from any particular library. They now consist of paid staff as well as volunteers from not one but multiple libraries. Policies and procedures have become increasingly detailed and complex.

Collection Development Policies and Criteria

Portals have created, developed, and refined specific collection development policies and selection criteria. This information may be available on their sites. Site selectors and reviewers often have access to additional and even more detailed guidelines and criteria.

Standardization

Site Design, Record Content, Indexing, and Abstracting

Overall design of portal sites is becoming more uniform. Initial screens usually display top hierarchical subjects and a search box. Simple and advanced search screens are available in most portals. Increasingly, they resemble the interfaces of subscription databases.

Table 7.

Name of Web Rating and Evaluation Site:	Academic Info
Site URL:	http://www.academicinfo.net/
Mission Statement, Description, Audience:	"To provide students, educators, and librarians with an easy to use online subject directory to access quality, relevant, and current Internet resources on each academic discipline" (cf. http://www.academicinfo.net/). Focus is on students in high school and above.
Year Founded:	1998
Origins/History:	Began as a for-profit site. In 2002, it was registered in the State of Washington as a non-profit organization.
Approximate Number of Records:	Not found on site
Selection Criteria:	Specific collection development policy with criteria is available on-site, currently at http://www.academicinfo.net/cdp.html.
Annotations?	Mostly quotes from sites themselves
Sites Rated? (e.g., with graphics such as stars)	NO
Browsable?	YES, by hierarchical classification, general to specific
Searchable?	YES, by keyword only. Boolean operators supported Default operator is "or."
Classification System Used?	NO
Subject Headings/Controlled Vocabulary?	NO
E-Mail Announcements/ Alerts for New Sites Added?	YES, monthly list.
Staffing:	Mike Madin, President of Academic Info
Responsible Person(s)/ Institution(s):	Mike Madin
Funding and Support:	Corporate and individual sponsors
Hosted by:	Site has its own server
Prime URL for "about" information:	http://www.academicinfo.net/cdp.html
COMMENTS:	"Academic Info relies on donations and sponsors to fulfill its mission."

Indexable Fields

Standardization of indexable fields in database records will allow portals to exchange information more freely and, if Z39.50 compliant, to facilitate searches across multiple portals. Standardization is important whether existing portals merge to form a single large database resource or whether they continue to exist separately.

Bibliographic Content and Control

Enhanced Record Content with Multiple Access Points

Increasingly, database records include distinct fields that provide multiple access points including personal or corporate author, title, description, subject headings, and in some cases even classification numbers (usually Dewey or LC).

Sophisticated Search Features

Most portals now offer sophisticated browse and search capabilities. Increasingly, complex searches are available utilizing Boolean operators, phrase searching, truncation, and more. Previously, such features were found primarily in subscription databases.

Interactivity

Features now commonly available—including e-mail alerts, comment and feedback buttons, and forms to suggest resources for inclusion—allow users to both contribute to and provide feedback to portals.

Cooperation

Recruitment of Libraries and Librarian Contributors

Some portal sites, such as Infomine, are actively recruiting libraries and librarians to contribute records. This is an extension of interactivity, noted earlier.

Current Issues and Future Trends

Many library groups and professional associations including the Library of Congress, Association of Research Libraries, the American Library Association's LITA Internet Portals Interest Group, and "Libraries of Organized Online Knowledge" (or LOOK, formerly FIAT LUX) are actively involved in encouraging and sponsoring research and planning for future portal development (cf. Library of Congress, 2003; American Library Association Library and Information Technology Association, n.d.; Association of Research Libraries 2003; Infomine, n.d.).

Mary E. Jackson, ARL Senior Program Office for Access Services, describes an intriguing vision of a "dream portal":

> Imagine one web site that can combine the powerful searching of web resources with the searching of local catalogs, online journals, or locally digitized resources. Add to this the ability to initiate a reference question, submit an interlibrary loan (ILL) request, and transfer into course management systems a citation or portion of a journal article, all without leaving that web site. (Jackson, 2002)

Jackson also shares the vision of Sarah Michalak, director of the University of Utah Libraries and a member of the ARL Scholars Portal Working Group, of a dream portal as

> a super discovery tool that specializes in high-quality content. The dream portal is fast and powerful. It searches across formats and resources and returns results that are deduped and relevancy ranked. It is more than a discovery tool because it delivers full text or information objects whenever available. The dream portal integrates appropriate applications such as course management software. Finally, the dream portal supports authentication and permits customization and personalization, e.g., alerts, saved hits or searches, and custom views of resources. (Jackson, 2002)

Key elements in these visions include a single point of access to high quality resources and databases (something commercial search engines and portals are less equipped to offer), integration of information in multiple formats, integration with other portals and software, interactivity including access to library services such as reference and interlibrary loan, provision of full-text whenever possible, and customization by users.

Towards these ends ARL, LC, LITA/IPIG, and other groups are developing or promoting "best practices," standards, cooperative projects, and sophisticated software to aid libraries and library groups in creating their own portals. They have met at ALA conferences and hope to chart the future course of librarian-created portals. Additional trends are noted and discussed below.

Content Access to Content Production

Initially, portals sought to index resources available externally. Many portals now either produce their own content or make content available on site. These include Internet Public Library and MEL. In the case of MEL, it provides significant amounts of copyrighted materials available only to Michigan constituents and so now are also, in a sense, subscription databases. Some, like IPL, MEL, and BUBL are now virtual libraries in addition to portals. This trend will continue.

Single Portal or Multiple Portals

Mason (2000) outlines several possible future directions for portals. Choices that are yet to be made include whether or not portals will merge

into a single resource or whether they will continue existing separately with increased cooperation and even interconnectivity. However, efforts by ARL, LC, and LITA definitely point not only to continuation of interconnected multiple portals but even to creation of new ones.

Resource Sharing

Some portals (notably Infomine) have developed open software made available to libraries and consortia who may wish to create their own portals. LC lists vendors of portal software on its Web site (Library of Congress, 2003). Some, like MEL, are considering making broad-based non—Michigan oriented content available to regional MELs, which would then provide their own local content.

Full-Text Capture

In an article about Librarians' Index to the Internet, Buchwald (2002) talks about LII and by implication other portals being able to "have some type of crawler like a regular search engine . . . [which] would need to capture the text of the selected homepage, and any meta tags and other keywords to build a useful fulltext index." These may include invisible information added to Web pages using "the Dublin Core, a means of building catalogue information into Web pages by using metatags, labels which exist in the unseen 'head' area of every online page" (Ansdell, 2000). Buchwald (2002) points out this may be more difficult, "since more and more, university, library, and newspaper sites are having areas of their sites blocked off from search engines' robots and crawlers." If such information could be captured, it would allow for more precise indexing, searching, and retrieval of Internet resources.

Broad vs. Highly Selective Resource Coverage

Infomine is seeking more comprehensive coverage of resources while BUBL:Link focuses more on including fewer yet highly selective resources (Dawson, 1997, p.18).

Conclusion

In less than a decade, librarian-created portals have changed dramatically in terms of growth, content, accessibility, interactivity, and organization. Many serve as virtual libraries, in some cases providing copyrighted content like subscription databases to specific clientele/constituents. Some have focused on substantially increasing resource coverage to compete more with commercial directories and search engines while others are less focused on growth and more on highly relevant resources.

Major issues include:

- Single, cooperatively produced and maintained portal vs. multiple portals increasing their interconnectivity and standardizing their content;

- Dramatic increase in the number of resources vs. a limited number of resources but of high quality;
- Development and sharing of sophisticated software to find, select, evaluate, index, and describe Web content as well as to provide bibliographic control within portals (cf. Schneider 2002a);
- Cooperative efforts to fund portal development (cf. Schneider 2002a);
- Increased efforts to globalize content.

As Schneider (2002a) aptly states: "We aren't going to blow the commercial portals out of the water. But we can be to the Internet what public radio and television are for these other media: a single place for local and global content that our public can trust."

REFERENCES

American Library Association. Library and Information Technology Association. (n.d.). *LITA Internet Portals Interest Group.* Retrieved May 14, 2003, from http://litaipig.ucr.edu/.

American Library Association. Library and Information Technology Association. (2003). *Internet Portals Interest Group.* Retrieved May 14, 2003, from http://www.ala.org/Content/NavigationMenu/LITA/LITA_Membership/LITA_Interest_Groups/Internet_Portals/Internet_Portals.htm.

Ansdell, I. (2000). Something to shout about in library. *The Herald.* [n.p.] Retrieved December 13, 2002, from http://www.theherald.co.uk/icon/archive/28-7-19100-22-0-11.html.

Association of Research Libraries. (2003). *ARL Scholars Portal Working Group.* Retrieved May 16, 2003, from http://www.arl.org/access/scholarsportal/index.html.

Buchwald, N. (2002). Standard review: Librarians' Index to the Internet. *Charleston Advisor, 4*(2), 37–41. Retrieved November 25, 2002, from http://charlestonco.com/dnloads/v4n2.pdf.

Dawson, A. (1997). BUBL bursts out of bath. *Serials Librarian, 31*(4), 15ff. Retrieved December 13, 2002, from http://search.epnet.com/direct.asp?an=9710034886&db=afh.

Infomine. (n.d.). *Overview.* Retrieved February 21, 2003, from http://infomine.ucr.edu/projects/.

Jackson, M. E. (2002). The advent of portals. *Library Journal, 127*(15), 36–39. Retrieved May 15, 2003, from Academic Search Elite database at http://search.epnet.com/direct.asp?an=7373456&db=afh. Also available free at http://libraryjournal.reviewsnews.com/index.asp?layout=articleArchive&articleid=CA242296%20&publication=libraryjournal.

Lawrence, S., & Giles, C. L. (n.d.). Accessibility and distribution of information on the Web. [Summary of Lawrence & Giles article from 8 July 1999 issue of *Nature*]. Retrieved April 2, 2003, from http://wwwmetrics.com/.

Library of Congress. (2003). *The Library of Congress Portals Applications Issues Group home page.* Retrieved May 15, 2003, from http://www.loc.gov/catdir/lcpaig/ paig.html.

Mason, J., et al. (2000). Infomine: Promising directions in virtual library development. *First Monday, 5*(6). Retrieved February 21, 2003, from http://www.dlib.org/dlib/january03/mitchell/01mitchell.html.

Mitchell, S., et al. (2003). iVia Open Source Virtual Library System. *D-Lib Magazine 9*(1). Retrieved February 21, 2003, from http://www.dlib.org/dlib/january03/mitchell/01mitchell.html.

Schneider, K. G. (2002a). Creating a Yahoo! with values. *Library Journal Net Connect, 127*(12), 36–37. Retrieved November 25, 2002, from Academic Search Elite database at http://search.epnet.com/direct.asp?an=6976758&db=afh.

Schneider, K. G. (2002b). Fiat Lux: A Yahoo with values and a brain. *American Libraries, 33*(4), 92. Retrieved November 18, 2002, from Academic Search Elite database at http://search.epnet.com/direct.asp?an=6431511&db=afh.

Search Engines. (2001). *Online, 25*(4), 14. Retrieved January 14, 2003, from Academic Search Elite database at http://search.epnet.com/direct.asp?an=5010675&db=afh.

Wells, A. T., Calcari, S., & Koplow, T. (1999). *The amazing Internet challenge: How leading projects use library skills to organize the Web.* Chicago: American Library Association.

Williamson, A. P. (2000). BUBL LINK /5:15: Smarter than the average search engine. *Serials Librarian, 37*(4), 37ff. Retrieved December 13, 2002, from Academic Search Elite database at: http://search.epnet.com/direct.asp?an=3062565&db=afh.

Additional Readings

Carter, D. S. (2000). Featured collection: The Internet Public Library. *D-Lib Magazine.* Retrieved December 12, 2002, from http://www.dlib.org/dlib/january00/01featured-collection.html.

Dawson, A., & Simpson, J. (1997). BUBL: How BUBL benefits academic librarians. *Ariadne, 10.* Retrieved December 13, 2002, from http://www.ariadne.ac.uk/issue10/bubl/.

Diaz, K. R. (1999). Internet Scout Project. *Reference & User Services Quarterly, 38*(4), 347ff. Retrieved January 14, 2003, from Academic Search Elite at http://search.epnet.com/direct.asp?an=2401166&db=afh; ILL request submitted online January 14, 2003.

Jacsó, P. (2001, March). Librarians' index to the Internet. *Reference Reviews Archive.* Retrieved November 25, 2002, from http://www.galegroup.com/servlet/HTMLFileServlet?imprint=9999&region=7&fileName=reference/archive/200103/lii.html.

Jacsó, P. (2001, November/December). Peter's picks and pans. *Online, 25*(6), 84–88.

Kenney, B. (2002). Sue Davidsen named director of Internet Public Library. *Library Journal Net Connect, 127*(12), 7. Retrieved December 5, 2002, from Academic Search Elite database at http://search.epnet.com/direct.asp?an=6976681&db=afh.

Lawrence, S., & Giles, L. (1999). Accessibility and distribution of information on the Web. *Nature, 400*(6740), 107–109. Retrieved April 2, 2003, from Academic Search Elite database at http://search.epnet.com/direct.asp?an=2017319&db=afh.

Leita, C., & Hinman, H. (1998). A public librarian helps launch an index. *Library Journal, 123*(16), 48. Retrieved November 25, 2002, from Academic Search Elite database at http://search.epnet.com/direct.asp?an=1131017&db=afh.

Librarians' Index to the Internet: "By librarians, for everyone!" (2000). *Connection* (California State Library and Library of California), *4,* 1–2. Retrieved November 25, 2002, from http://www.library.ca.gov/newsletter/2000/CSL_Connection_Oct00.pdf.

Matthews, J., & Wiggins, R. (2001, May 15). MEL: A gardener's perspective. *Library Journal, 126*(9), 37. Retrieved November 25, 2002, from Academic Search Elite database at http://search.epnet.com/direct.asp?an=4454952&db=afh.

McDermott, I. E. (2000). Classified with class: Superior subject sites. *Searcher, 8*(4), 10, 12–14, 16, 18.

Mitchell, S., & Mooney, M. (1996). INFOMINE—A model Web-based academic virtual library. *Information Technology and Libraries, 15*(1), 20–25.

Nicholson, D. (1996). BUBL with a Z spells . . . LINK? The reinvention of *BUBL. Computers in Libraries, 16*(2), 82–83. Retrieved December 13, 2002, from http://search.epnet.com/ direct.asp?an=9602203624&db=afh.

Oder, N. (1998, October 1). Cataloging the net: Can we do it? *Library Journal, 123*(16), 47–51. Retrieved November 25, 2002, from Academic Search Elite database at http://search.epnet.com/direct.asp?an=1131011&db=afh.

Oder, N. (2000). Cataloging the net: Two years later. *Library Journal, 125*(16), 50–51. Retrieved January 3, 2003, from Academic Search Elite database at http://search.epnet.com/direct.asp?an=3637498&db=afh.

Paynet, G., et. al. (2002). The Infomine Project. [PowerPoint presentation given at ALA 2002]. Retrieved January 12, 2003, from http://infomine.ucr.edu/publications/ALA2002/HTML/img0.html.

Price, G. (1999). Internet Scout Project looks toward the future. *Information Today, 16*(11), 35. Retrieved January 14, 2003, from Academic Search Elite at http://search. epnet.com/direct.asp?an=2567184&db=afh.

Rogers, M., & Oder, N. (2001). Schneider to head librarians' index. *Library Journal, 126*(13), 24.

Wise, M. (1999). Academic info. *College & Research Libraries News, 60*(2). Retrieved November 18, 2002, from http://www.bowdoin.edu/~samato/IRA/reviews/issues/feb99/academic.html.

Access in a Networked World: Scholars Portal in Context

Jerry D. Campbell

Abstract

Since the 1960s, librarians have projected a vision of a digital library that offers seamless access to a vast world of scholarly information of all types. Until the 1990s, however, digital technologies lacked the power and capacity to deliver on the vision. During the 1990s, technology platforms, networking technologies, electronic resources, and the evolution of standards matured sufficiently to lay the foundation for the vision to be fulfilled. As this maturation was taking place, the rapid growth of electronic resources was based on numerous proprietary systems making access across such systems impossible. The scholars portal project is an effort to create a search and retrieval tool that will provide an interim solution to this problem until such time as those systems are built on a unified set of standards and data formats.

Since the publication of the Association of Research Libraries (ARL) white paper on the need for a research library portal (Campbell, 2000), the concept of a *scholars portal* (SP) has generated much interest. Illustrative of this interest are what might be described as several independent demonstration projects sponsored by a number of entities including individual libraries, ARL,[1] the Ontario Council of University Libraries ("What's New," 2002), and the Council of Australian University Librarians.[2]

Interest in an SP also generically characterizes a number of efforts to create specialized subject portals for researchers.[3] Initially described as "the place to start for anyone seeking academically sound information" (Campbell, 2000, p. 211), the SP concept has been widely explored, developed, and refined.[4]

Jerry D. Campbell, Dean, University Librarian, and Chief Information Officer, University of Southern California, Doheny Memorial Library, Los Angeles, CA 90089–0182
LIBRARY TRENDS, Vol. 52, No. 2, Fall 2003, pp. 247–255

The purpose of this essay, therefore, is not to reiterate the definition of the now well-published concept of an SP but rather to outline the larger context within which SP exists. It is often our tendency to place exaggerated expectations on new technologies and thereby diminish the value of their eventual impact. Alone, SP constitutes only one small but vital step in the much larger jigsaw puzzle of the evolving digital library. Thus, it is important to place SP in this larger context and to understand the nature of the contribution it will make to the advancement of digital libraries.

Grand Vision

Since the earliest days of digital technology, we have been sharing grand visions about how it would transform the things we do. These dreams, of course, have included speculations about how libraries might be changed. By their nature, such visions often do not deal with details but rather focus on the larger dream. After the first decade of serious speculation about technology and libraries,[5] one study suggested that technology would transform the structure and function of libraries of the 1980s by storing materials in new formats, by making "obsolete the concept of the catalog and the book stack" as they were then known, and by linking them by means of a nationwide network (Conference Board, 1972, pp. 116–117). One of the best recent statements of the vision is equally elegant in its simplicity and challenging in its scope: "The dream to which we need to aspire is that all scholarly and research publications (including university, governmental, research, and museum sites) be universally available on the Internet in perpetuity" (Hawkins, 2000). Taken together, such recurring projections of the technology empowered library of the future have kept before us the vision of a slowly emerging digital library.

At almost any point over the past four decades, however, the challenge of these visions has outstripped the actual capacity of digital technology to deliver the dream. With hindsight, we can see that over the years virtually every aspect of the technology, from power and capacity to programming language, has been unequal to the challenge. Absent also have been the required infrastructures of connectivity and data standards necessary for the dreamed digital library to function. We now know that an operating digital library requires a vast number of elements functioning flawlessly together. Glimpsing the vision, it turns out, has been far easier than bringing it to reality.

All Things Necessary

Only recently have we begun to have in prospect all things necessary to implement the library envisioned for a generation. Among the vast number of improvements in computer technology, the following categories are key for the development of digital libraries.

Perfecting the Platforms

Sometime during the 1990s the persistence of the Moore's Law phenomenon produced computing hardware platforms of sufficient speed and memory to begin to implement the long-articulated vision. Amidst constantly improving machine specifications, it is not possible to isolate the moment it happened. Indeed, it is not possible to articulate exactly what speed and memory capacity were necessary. It is only possible to look back and recognize that during that decade we began to have access to hardware platforms of sufficient capacities to develop functional, if still rudimentary, digital libraries.

Similarly, these capable computers were accompanied by the development of other necessary components. Among these were

- the evolution of online storage systems with the capacity to hold and make available quickly massive amounts of information pertinent to digital libraries;
- the availability of increasingly sophisticated programming languages and software platforms;
- the introduction of improved systems for authentication and authorization.

This is not to argue that the computer platforms reached an end point or even slowed in their evolution in the 1990s. To the contrary, Moore's Law continues unabated (Kurzweil, 1999, pp. 20–25). It is only that they achieved sufficient capacities to allow implementation of digital libraries to begin in earnest.

Achieving the Connectivity

Even with the early Internet at our disposal, the possibility of highly networked libraries could not be realized as the 1990s arrived. Unlike with platforms, however, it is clear when a world-changing advance in connectivity occurred. In 1993 Tim Berners-Lee introduced the World Wide Web (the Web), which provided the infrastructure for flexible use of the Internet, thereby transforming connectivity (Berners-Lee & Fischetti, 1999). The Web rapidly became the mechanism by which libraries (and everything else) sought to achieve giant strides in the networking of resources.

Soon thereafter, the Internet itself was challenged to meet the growing bandwidth requirements stimulated by the Web as the so-called "commodity" Internet was born. As a consequence, Internet2 (I2) was launched to provide much greater bandwidths in a separate network environment for the research community. I2 technology made possible the rapid exchange of large files and empowered libraries to move from text to larger files such as those containing graphic materials.

As with platforms, the technology of connectivity continues to improve. Efforts are already underway to develop even higher speed optical networks. In addition, wireless data networks have recently added much needed nomadic flexibility for network users. Though wireless networks are comparatively slow shared environments, they, too, are rapidly increasing in bandwidth.

Evolving a Critical Mass of E-Resources

The 1990s also saw an amazing growth in the amount of information available on the Web (Lyman & Varian, 2000, p. 5). The surge in online information included journal literature from both for-profit and not-for-profit publishers as well as a rich sampling of archival resources principally from universities and national libraries. In addition, the decade brought a phenomenal surge in the production of raw data from the growth of computational science. In the panoply of published information, only monographic literature lagged behind, primarily because monographic publishers were slow to embrace the technology and because copyright restrictions served as an impediment (Lynch, 2002). All in all, therefore, the 1990s witnessed an extraordinary growth in the availability of digital information.

Creating Containers and Hooks

Unfortunately, the rapid growth of information available in digital format was marked by a significant problem from the standpoint of library users: it was characterized by a multitude of formats that did not offer a uniform means by which it might be found. In other words, the rapid growth of digital information began before we had developed common standards for data and metadata. As a result, users have not had an easy way to identify and retrieve information with thoroughness and precision. In this situation, users, especially neophytes, have had a difficult time identifying pertinent information, and thorough research has often been difficult even for experienced users. In addition, as the wealth of digital information continued to grow, the problem only worsened because the variety of formats and lack of metadata persisted.

Thus, the 1990s saw significant efforts to address the need for common formats for data and metadata. Centering on the Web environment, a number of data formats were introduced, chief among them being SGML (on which HTML is based) and more recently XML (Berners-Lee, 2002). Similarly, a number of metadata formats were developed, including Dublin Core, EAD, METS, and MODS (Tennent, 2002). These efforts, however, are ongoing and cannot be said to be fully mature. Thus, while they have had a positive impact by slowing the proliferation of data and metadata variations, they have not fully resolved the problems associated

with what might be described as islands of disconnected data and information.

The Rise of Search Engines

With the early rise of information contained in data sources (databases and data repositories) in the 1960s came the need to provide tools for its retrieval. Such tools were first developed in connection with specific data sources. To function, information or objects in a data source first had to be described or indexed in a manner that could be interpreted by machine. These are the descriptions we now generally refer to as metadata. A "search engine" was then developed and customized to the indexing scheme, making it possible to retrieve data. Depending upon the quality of the indexing and search engine, such tools allowed users to find and retrieve specific data objects from data sources with speed and precision. Most significant data sources were accompanied by customized, unique search engines.

Later, as database architecture became better developed, a number of major providers offered the capacity to search across large amounts of information as long as it was stored in their proprietary syntax for expressing structure in data. This improved the general situation, but still left it difficult to work with information spread among different proprietary solutions. The economics of information and the mechanism of copyright law have provided significant disincentives for solving this problem of standalone data sources.

With the introduction of the Web and the vast amount of information located there (or, perhaps, lost there) came the urgent need to develop tools for retrieval in the Web environment. Thus, in the mid-1990s, a number of agencies developed search engines designed specifically for the Web. These engines, commonly referred to as portals, primarily searched for keywords and phrases and applied relevancy ranking schemes to determine validity. They also developed their own indexes from keywords and phrases in order to carry out searches quickly. As they have improved, they have become sophisticated, powerful, and amazingly adept at locating information on the Internet.

The sophistication and power of evolving search engines notwithstanding, however, certain limitations still confronted library users. The data sources served by specific search engines were still a disconnected sea of information islands whose contents could not be discovered and retrieved by a single search tool. These unique search engines characterized most digital information and data licensed by libraries. Furthermore, the extraordinary Web search engines could not see the actual data beneath such search engine–driven data sources even when those data sources were Web enabled because the Web only provided access to the

unique search engines themselves. Indeed, the Web itself became another information island (albeit from the standpoint of size, continent might be a more accurate metaphor) and compounded the problem for those seeking information.

Scholars Portal

It was with respect to the context outlined above that the concept of an SP was developed. The growing plethora of high-quality E-resources created a clear need for a tool that could, with a single search interface with a multitude of unique search engines (including Web engines), discover and retrieve relevant information from each, and present a single, merged set of results to the researcher. While other desirable features have been included in the Association of Research Libraries SP Project, this is its fundamental purpose and the need that drove its conceptualization there and elsewhere. The combination of capable platforms, high bandwidth connectivity, maturing data and metadata formats, and sophisticated but target-limited search engines was sufficient to indicate that such a portal was possible and, moreover, suggested where it should fit in the panoply of information technology.

Specifically, an SP would serve as an aggregator of search engine–driven data sources. Initially, it would necessarily be limited in scope because it would require that interfaces be created (programmed) and kept up-to-date for each data source. In their first implementations, therefore, SPs will be institution- or agency-specific with a defined set of data sources identified as targets to be aggregated. They might be thought of as offering second-level search engines in that an SP engine would sit above other search engines. In many cases, the nature of license restrictions on data sources suggests this approach. Even when local SPs become linked via the Web, certain limitations on the retrieval of information may be required at each location, based on the number and nature of institution- or agency-specific licenses. Nonetheless, by searching across even a limited set of data sources, SPs will vastly improve the prospect that digital library users will be able to discover and retrieve high-quality information.

Glimpsing the Future

As much as they are needed, SPs constitute a poor solution to a complex problem. They are poor solutions because they represent yet another layer of technology necessary to solve problems created by earlier layers of technology and because they must be adjusted each time underlying technology changes. Thus, SPs may best be thought of as an interim but necessary step in the evolution of tomorrow's digital library, a step that will be made obsolete upon the eventual emergence and utilization of accepted standards for data and metadata along with a new generation of tools for

searching. Such standards and new tools could both reduce the number of technology layers and increase the ease of information discovery and retrieval.

It should be noted that the perfection of standards and new discovery and retrieval tools will not alone alter the impact of economic incentives for providers to continue to maintain separate data sources. Only if the habits of scholars come to express a preference for standards-based sources and tools to the exclusion of separate data sources will the economic incentives be reversed.

Such standards and tools are only now being developed and will be some time in development. Until the Web is supplanted, they will also be Web-based technologies. In addition to the continued evolution of the standards for data and metadata noted above, an example of a new approach is the OpenURL effort (Stern, 2001). This solution would function by running data mining processes (so-called "smart agents") against a generic, public syntax (OpenURL) resource identification system, by utilizing an identified local "resolver" machine to validate and give the location of items, and by employing extensible metadata syntax to standardize and store query data from a variety of metadata formats. While such a system is feasible, as the previous sentence indicates, it is complex and requires not only agreement on the OpenURL syntax but the development of viable smart agents and "resolver" machines loaded with appropriate data. It will also require time.

Indeed, given the current situation and in spite of our considerable technological prowess, no ultimate solution to the fragmentation of data sources is likely to be simple or quick in its development. And compared to any ultimate solution, scholars portals are much simpler and already available in first-generation versions. This indicates that efforts to test scholars portals, even if only as interim solutions to the problem, are necessary and justifiable. Not only will they move us a little closer to the dream of a universal, networked digital library, they will also give our users something they urgently need today.

As for the grand vision of the digital library of the future, it will eventually come to pass. In time, "all scholarly and research publications (including university, governmental, research, and museum sites)" will indeed "be universally available on the Internet in perpetuity" (Hawkins, 2000). It may be hard today to believe that such an outcome will be achieved, but a scant decade ago it would have been equally hard to believe that something called "the Web" would transform not only the distribution of knowledge but the habits of the workplace as well. It is important, therefore, that we continue to believe in the vision and that we continue to articulate it. It is also important that we work to make it a reality.

NOTES

1. See "Seven ARL Libraries" (2002) and Quint (2002).
2. See http://library.queensu.ca/libguides/databases/scholarsportal.htm; http://anu.edu.au/caul/caul-doc/caul20022aarlin.doc.
3. See Halbert (2002). See also Technical Issues ad hoc Committee (2002).
4. See the substantive article by the ARL Portal Project manager M. E. Jackson (2002). See also Thomas (2000a).
5. See the summary in Fussler (1973, pp. 1–11).

REFERENCES

Berners-Lee, T. (2002). *W3C data formats.* Retrieved December 2, 2002, from http://www.w3.org/TR/NOTE-rdfarch.html.

Berners-Lee, T., & Fischetti, M. (1999). *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor,* 1st ed. New York: HarperCollins.

Brown, J. S. (2002). Learning in the digital age. In *Futures Forum 2002: Exploring the future of higher education* (pp. 20–23). Cambridge, MA: Forum for the Future of Higher Education.

Campbell, J. D. (2000). The case for creating a scholars portal to the Web: A white paper. *ARL Newsletter 211.*

Conference Board. (1972). *Information technology: Some critical implications for decision makers.* New York, NY: Conference Board, Inc.

Fussler, H. H. (1973). *Research libraries and technology.* Chicago: University of Chicago Press.

Fyffe, R. (2002). Technological change and the scholarly Communications Reform Movement. *Library Resources & Technical Services, 45*(2), 50–61.

Gillespie, R. G., & Dicaro, D. A. (1981). *Computing and higher education: An accidental revolution.* Seattle: University of Washington Press.

Greenstein, D., & Thorin, S. E. (2002). *The digital library: A biography.* Washington, DC: Digital Library Federation and Council on Library and Information Resources.

Halbert, M. (2002). The Metaarchive.org Project: A joint project of Emory University and ASERL. Retrieved August 26, 2002, from http://www.metascholar.org/modules.php?op=modload&name=News&file=article&sid=4.

Hawkins, B. (2000). Libraries, knowledge management, and higher education in an electronic environment. Retrieved October 7, 2002, from http://www.alia.org.au/conferences/alia2000/proceedings/brian.hawkins.html.

Herring, S. D. (2002). Use of electronic resources in scholarly electronic journals: A citation analysis. *College & Research Libraries, 63*(4), 334–340. Retrieved June 23, 2003, from http://library.queensu.ca/libguides/databases/scholarsportal.htm.

Jackson, M. E. (2002). The advent of portals. *Library Journal.* Retrieved October 28, 2002, from http://libraryjournal.reviewsnews.com/index.asp?layout=articlePrint&articleIDCA242296.

Kurzweil, R. (1999). *The age of spiritual machines: When computers exceed human intelligence.* New York: Penguin Books.

Lougee, W. P. (2002). *Diffuse libraries: Emergent roles for the research library in the digital age.* Washington, DC: Council on Library and Information Resources.

Lyman, P., & Varian, H. (2000). How much information? Retrieved August 8, 2002, from http://www.sims.berkeley.edu/research/projects/how-much-info/summary.html, p. 5.

Lynch, C. (2002). What's become of the digital library? Retrieved December 4, 2002, from http://www.educause.edu/asp/doclib/abstract.asp?ID=EDU0207.

NetLibrary and eBooks an excellent fit for OCLC. (2002). *OCLC Newsletter, 256,* 30–31.

Project MARC Reports. (1968). *Library Resources and Technical Services, 12*(3), 245–319.

Quint, B. (2002). Academic libraries develop integrated portal software package. In *Newsbreaks.* Retrieved December 4, 2002, from http://www.infotoday.com/newsbreaks/nb020513-2.htm.

Seven ARL libraries launch Scholars Portal Project. (2002). *Library Journal.* Retrieved December 4, 2002, from http://libraryjournal.reviewsnews.com/index.asp?layout=articlePrint&articleID=CA215696.

Stern, D. (2001). Automating enhanced discovery and delivery: The OpenURL possibilities. *Online.* Retrieved August 29, 2002, from http://www.infotoday.com/online/OL2001/stern3_01.html.

Technical Issues ad hoc Committee on Content Linking and Management for Electronic Serials and other Digital Materials. (2002). Working bibliography. Retrieved August 28, 2002, from http://www.viva.lib.va.us/viva/tech/cat/link/working_bibliography.html.

Tennant, R. (2002). Digital libraries—Metadata as if libraries depended on it. Retrieved August 26, 2002, from http://libraryjournal.reviewsnews.com/index.asp?layout=articleID=CA206408&.

Thomas, S. E. (2000). Abundance, attention, and access: Of portals and catalogs. Retrieved December 6, 2002, from http://www.arl.org/newsltr/212/portal.html.

Thomas, S. E. (2000). The catalog as portal to the Internet. Retrieved August 26, 2002, from http://lcweb.loc.gov/catdir/bibcontrol/thomas_paper.html.

Travis, T. A., & Norlin, E. (2002). Testing the competition: Usability of commercial information sites compared to academic library Web sites. *College & Research Libraries, 63*(5) 433–448.

What's new: Ontario Scholars Portal. (2002). Retrieved December 4, 2002, from http://www.uottawa.ca/library/actnew-e.html.

Government Information on the Internet

Greg R. Notess

Abstract

The U.S. federal government has been a major publisher on the Internet. Its many agencies have used the Internet, and the Web most recently, to provide access to a great quantity of their information output. Several agencies such as the Library of Congress and the Government Printing Office have taken a leading role in both providing information and offering finding aids, while other endeavors such as FirstGov and subject gateways offer other avenues of access. A brief look back at the history of the government on the Web and the continuing concerns and challenges show how the government is not only a major content provider on the Internet but also a source for the organization of the content.

Introduction: Government Information Dissemination and the Internet

The United States federal government produces a great quantity of information and has been one of the largest publishers in the world. Throughout the twentieth century the amount of information from the federal government has increased enormously. Consider just the number of physical volumes published for each of the decennial censuses and how with each census until the most recent the number of print volumes has grown tremendously. The rest of the government's corpus increased in a similar fashion.

The ideal of the free flow of government information to the people grew into the Federal Depository Library Program (FDLP), with libraries in every state as a means to achieve that ideal. The FDLP certainly provided

Greg R. Notess, Reference Librarian and Professor, Montana State University, Bozeman, P.O. Box 173320, MSU Library, Bozeman, MT 59717–3320
LIBRARY TRENDS, Vol. 52, No. 2, Fall 2003, pp. 256–257

unprecedented access to government documents to a significant portion of the country's citizens. In the latter part of twentieth century, the Paperwork Reduction Act and other legislative and regulatory efforts showed a significant concern with the cost to the government of both the printing of all of these documents and the expense of disseminating them to so many libraries.

Meanwhile, another government effort, ARPANET from the Advanced Research Projects Agency (ARPA, later known as the Defense Advanced Research Projects Agency or DARPA), was creating the beginnings of the Internet. Cold War fears led the researchers to look into using packet switching technology for the network to survive nuclear bombing attacks taking out large sections of the network.

As the Internet developed, the information dissemination capabilities of a large network became apparent to those involved in the research. Network developers used it themselves to communicate with each other, and electronic mail became one of the principal means of electronic communications. In addition to brief messages, researchers began sharing documents and then databases. Basically, the Internet became a way to share information.

The federal government certainly has made great use of the Internet for the dissemination and organization of its publications. Since the early days of the Internet, government information resources have grown and expanded in scope. From the Library of Congress to the Government Printing Office and many others, a great quantity of information content has been made available online, and a variety of finding aids and search engines help provide access. While there remain gaps, concerns, and challenges, the government is a major provider of quality, substantial content on the Internet.

Library of Congress

The Library of Congress (LC) has been organizing print resources for decades. The Library of Congress Classification System and Subject Headings are staples of library organization. While they have not brought the same level of organization to the Internet, they have certainly contributed some major resources. A look back at the brief history of LC on the Internet provides an example on a large scale of what many other government agencies have done.

Start with the April 30, 1993, announcement from LC. On that date, the Library announced its accessibility on the Internet when it made the Library of Congress Information System (LOCIS) available via telnet connections. LC had joined the hundreds of other libraries who freely offer their catalogs via the Internet.

Library catalogs occupy a unique role in the growth of information resources on the Internet. Internet availability on college campuses and at

government research labs in the 1980s meant that telnet was widely available and that it created new possibilities for information dissemination. And recently automated libraries had freely available online databases that they were happy to share. Lynch and Preston (1990) note that, by 1989, the Colorado Association of Research Libraries (CARL) catalogs and the University of California system catalog (MELVYL) were available on the Internet via the telnet protocol. The number of library catalogs rapidly expanded from there so that, by October 16, 1992, Billy Barron (1992) offered a listing of 482 Internet-accessible library catalogs.

But with the 1993 LC launch, the largest government library provided not only its catalog but a collection of other important federal information. Notess (1993) describes the various databases that LC made available via telnet, including a database of copyright registrations and another with information about federal legislation. This became one of the first free sources for information on federal proposed and passed legislation. The database was not just for current bills. It even covered legislation back to 1973.

The Library of Congress' offering of LOCIS, even back in 1993, demonstrates several trends and approaches to presenting and organizing government information on the Internet. LC data was available online even before LOCIS, but when LC opened up LOCIS to Internet users, it went well beyond descriptive agency information or agency-specific databases.

First of all, consider the LC catalog itself. Much of the data within the LC catalog had long been accessible online. Fee-based bibliographic utilities such as OCLC, RLIN, and WLN offered LC cataloging records to subscribers. And all three utilities were Internet-accessible by 1993.

More significant for the general Internet user was the Digital Research Associates' (DRA) service, which was often referred to as the "LC catalog." DRA was a library automation vendor, and before the Library of Congress itself opened up LOCIS via the Internet, DRA provided telnet access (originally at dra.com and still available at lcmarc.dra.com/lcmarc) to the LC-MARC bibliographic file and authority file (Rogers, 1992).

Before 1993 was over, LC moved on from the telnet-based LOCIS to using the newer menu-driven gopher technology and introduced LC MARVEL (Library of Congress Machine-Assisted Realization of the Virtual Electronic Library). MARVEL prefigured much of the kind of information content that most government agencies put online even today. It included sections for information about LC and MARVEL, links to LOCIS, press releases, library hours, information on how to obtain an ISBN or ISSN, and congressional information ("Library of Congress Goes Online," 1993).

The next year, LC announced plans to digitize some of its collections, such as photographs, maps, pamphlets, and speeches. It then planned to

make them accessible on the Internet and in particular wanted to promote education by making the collections accessible at schools and local libraries (DeLoughry, 1994). The great success of the American Memory Project (http://memory.loc.gov) grew out of this laudable goal.

By the beginning of 1995, LC had also announced the launch of its new Congressional Web site, THOMAS, at thomas.loc.gov/ (Library of Congress, 1995). THOMAS brought the full text of legislation, the House calendar, summaries of floor proceedings, and additional resources to the growing Web-using public. This move from telnet to gopher to the Web was mirrored by many other government agencies.

LC has come a long way since 1995, and its Web site is a gateway to THOMAS, American Memory, the LC Catalog, and much more. Yet this brief overview of its early history on the Internet highlights several trends seen elsewhere in online government resources including: that information is first put online by a nongovernment agency, the use of new technologies to disseminate information, and the digitization of documents.

Push from the Outside

Some of the other major government information resources were likewise first made available through the efforts and servers of nongovernmental entities before the government itself made the move.

The Internet Town Hall, a nongovernmental Internet site, was an active proponent of making the Security and Exchange Commission's (SEC) Electronic Data Gathering, Analysis, and Retrieval (EDGAR) filings available for free to the public. And while the SEC was busy pushing companies to file via EDGAR, thus ensuring that the SEC would have the information in electronic format, the Internet Town Hall influenced the SEC to then make the filings available to the public (Notess, 1994). Eventually, the SEC developed its own system and process, so that it now makes all the EDGAR filings available on its Web site.

Full-text U.S. patents are another well-known example. Several other nongovernmental companies also offered free access to full-text patents and related patent information (Santo, 1995). Two private organizations, MicroPatent and Source Translation and Optimization, offered some very useful free search services with access to patent abstracts and even some full-text patents. The Internet Town Hall also offered an experimental Internet publication of patent information in 1994. It provided free access to the full-text of recent patents. Shortly before their experiment ended officially, the Patent and Trademark Office (PTO) finally announced (September 26, 1995) that they would provide patent information on their Web site and that it would be available for free to the public. Yet Kaiser (1998) reports that the PTO was just announcing that the data would not be available until late in 1998 and early in 1999.

Government Printing Office

As the LC example showed, government agencies have continued to use new technologies to disseminate, organize, and present their information. The Government Printing Office (GPO) went through a similar transformation. But it had another factor affecting its changes.

The rise of the Internet and its incredible transformation in the 1990s from a technology experiment of limited interest to becoming one of the primary means of disseminating information occurred at the same time as a growing concern over government expenditures. The timing of the Internet's growth coincided with efforts to reduce the expense for the government of printing and disseminating publications and the expense of producing them. So the GPO and the Superintendent of Documents had great incentive to explore various options for transforming government publications into electronically disseminated documents.

Indeed, GPO has been in the forefront of government agencies in effectively using the Internet as a dissemination medium and has made significantly more efforts at providing bibliographic control of its output than many other government agencies. The recently relaunched GPO Access (now at http://www.gpoaccess.gov) has been one of the major sources of government information on the Internet for a decade. It also provides several important tools for the organization of online government information from other agencies.

GPO first began electronic dissemination by sending floppy disks and CD-ROMs to depository libraries in the 1980s. However, disks still share the same production and expense problems that print sources face. Multiple copies of each disk are produced and then sent to the depository libraries. Another approach from the 1980s was using an electronic bulletin board (BBS) for dissemination. Data could be produced just once, placed online at the BBS, and then users would dial into the BBS via modem (and perhaps long-distance charges) to retrieve the data. Yet most BBS interfaces were not easy to use and retrieving specific data could get quite complex. The long-distance phone charge also discouraged most general interest use.

With the rise in access to the Internet by the public and libraries, GPO then moved on to the Internet. Originating from the GPO Access Act, or more officially, the Government Printing Office Electronic Information Access Enhancement Act of 1993 (Public Law 103–40, 107 Stat. 112, June 8, 1993), GPO created publicly accessible and searchable access to major government publications like the *Congressional Record* and the *Federal Register*. As Minahan (1994) notes, GPO had the documents up on the Internet by the following year. Searchable access relied on Wide Area Information Service (WAIS), a sophisticated and free full-text search system with relevancy ranking developed by Brewster Kahle (1992) at Thinking Machines Corporation.

Although the law allowed a charge for access (except at depository libraries), the GPO Access databases soon became freely available to all through GPO Access Gateway sites set up by depository libraries ("GPO Access," 1995), and then by December 1995 GPO decided to waive the fees for GPO Access (Gordon-Murnane, 1999).

As of April 2003, GPO Access lists over ninety distinct databases, all accessible via the GPO Access site. Of those databases, several are particularly important in providing some level of bibliographic control over electronically published documents. The Catalog of U.S. Government Publications (http://www.gpoaccess.gov/cgp), the online successor to the venerable print *Monthly Catalog of U.S. Government Publications,* is a bibliographic catalog of print and electronic publications created by federal agencies from 1994 through the present.

The records for online documents include links to the online full-text publications when possible, but the GPO also provides another database, New Electronic Titles (http://www.access.gpo.gov/su_docs/locators/net), that focuses exclusively on Web-accessible federal government publications. Organized by month, New Electronic Titles actually does a specialized search in the Catalog of U.S. Government Publications for online documents.

The continued cataloging by the GPO of online documents, especially those that no longer have a print version, means that there is better descriptive and subject information about the documents than there is for most Web pages. GPO also decided to use Permanent Uniform Resource Locators (PURLs) rather than the more standard URLs.

The idea of PURLs is certainly worthy. URLs can and do change frequently. A document that used to be at http://agency.gov/latestreport.html may soon be moved to http://agency.gov/archive/crypticstringt.html. A GPO PURL will look more like purl.access.gpo.gov/GPO/LPS25. The PURL is then redirected to the appropriate URL. The permanence is achieved by having a PURL resolver that always has the updated URL. For libraries, this greatly eases record maintenance work. Only the PURL resolver needs to be updated, not every single library that has the record in their catalog.

On the negative side, PURLs do not provide the same level of information that a URL can. For example, the PURL for *Prague, NATO, and European Security* is http://purl.access.gpo.gov/GPO/LPS12869, which redirects to the URL of http://www.carlisle.army.mil/ssi/pubs/1996/prague/prague.pdf. A close look at the URL shows that this is a 1996 publication, from an army site, and it is in PDF format, none of which is obvious from the PURL. In addition, even the PURLs sometimes fail to have a functioning URL in the resolver database. At least the full record in the Catalog of U.S. Government Publications usually includes the URL as well as the PURL.

Governmental Site Organization

Many agencies went from BBS to gopher to the Web but, once they arrived on the Web there was still plenty of development and a concern with how best to present and organize the information on their site or sites. The U.S. federal government had been closely involved with the creation of the Internet, so it was natural that the government would also want to be a savvy user of the network. As government agencies began to make systems available via telnet, FTP, gopher, and eventually and most successfully on the Web, there was an obvious organizational structure already in place: the structure of the government

Thus, the typical first move online for a government agency has been to set up an agency Web page arranged hierarchically just like the agency. In the mid-1990s government departments set up Web pages that were often organized to mirror their internal structure. It certainly made sense to those within the agency, who knew the structure, but Web designers soon realized that it confused most other users.

Several alternative approaches have developed and, even though some Web sites are still primarily organized around the hierarchy of the agency, the government now has a wide diversity in the types of Web sites available.

One of the first approaches was to build a site organized by groups of users. The Library of Congress site (http://www.loc.gov) even has a section like this now, with separate entry points for Researchers, Law Researchers, Librarians & Archivists, Teachers, Kids & Families, Publishers, Persons with Disabilities, Blind Persons, and Newcomers. NASA's site (http://www.nasa.gov) highlights four target audience groups: Kids, Students, Educators, and Media and Press.

Another approach that started around 1998 was the construction of cross-agency, subject-specific sites. These subject-oriented gateways also had keyword-derived domain names rather than agency-related domains. For example, Healthfinder at http://www.healthfinder.gov was designed to assist consumers in finding government health information on the Internet. Recreation.gov offers information from all of the federal land management agencies that have recreational use on their lands. The U.S. Business Advisor (http://www.business.gov) aims to give businesses a central access point to government services, transactions, regulations, and opportunities.

The FEDSTAT site (http//www.fedstats.gov) follows this approach. It tries to provide quick and easy access to the broad range of statistics offered by more than one hundred federal agencies. With topic links, statistics by geography, and a multi–Web site search function, it offers several access points to the statistics. Yet, with the huge number of statistical reports covered, it can still be difficult to identify exactly the most pertinent report without having some knowledge of the whole universe of government-produced statistics.

Finding Government Internet Sources

The problem with the voluminous publications from the government and the corresponding voluminous number of Web sites is that, with so much information available, it can be quite difficult to know where to find an answer to a specific question. Fortunately, there are a number of finding aids and search tools that can make the task at least somewhat easier.

Many search and retrieval systems focused on U.S. government information resources have come and gone. In the early era of the Web, directories of government Web sites and gopher servers were the best entryways into the online government resources because they provided some kind of hierarchical agency access.

The government section of the well-known Yahoo! directory was updated relatively frequently in the early years and was fairly comprehensive. It also included sections at the federal, state, local, and international government levels. While useful for finding agencies, it did not work very well at finding government information by subject.

By 1997, the Federal Web Locator from the Villanova Center for Information Law and Policy was one of the most frequently used general directories of U.S. federal government agency Web sites. It was arranged by agency hierarchies that roughly mirrored the arrangement of the *U.S. Government Manual.* Although it did have a basic keyword search ability, it was still primarily a directory by agency rather than by subject. It has moved around between various URLs but can now be found at http://www.infoctr.edu/fwl.

Many other specialized directories have been created to help people find government Web sites. Almost all of them relied heavily on hierarchical agency access. At the same time, developments in search engines that would index all the words on Web pages and make them searchable were being developed and getting increasingly popular. On the government side of things, there were a few specialized search engines just for U.S. federal government sites.

GovBot, from the Center for Intelligent Information Retrieval (CIIR) at the University of Massachusetts, was one early example. GovBot used the Inquery software developed at the CIIR to search Web pages in the .gov and .mil domains. However, for some searches, general Web search engines such as HotBot or AltaVista using a .gov or .mil domain limit could be more effective. GovBot lasted for several years but was eventually retired.

Another interesting search engine solution was launched by Northern Light in the spring of 1999. USGOVSEARCH was the result of a partnership between the Web search engine Northern Light and the Commerce Department's National Technical Information Service (NTIS).

As Hane (1999) reports, USGOVSEARCH included over 20,000 U.S. government agency and military Web sites with almost 4 million pages,

along with the 2 million abstracts from NTIS. Because the initial May 1999 announcement included access pricing details, the search engine met with several protests about the pricing structure.

Yet, USGOVSEARCH offered some real advantages and features that differentiated it from other government-oriented search engines. It included the full NTIS abstracts database, which was available nowhere else on the Web for free. The advanced search included the subject limits derived from a Northern Light–created, government-oriented taxonomy. While it only listed the more general upper-level hierarchical terms, the more specific subject terms would be found in the folders within search results (Notess, 2000).

Unfortunately, the deal with NTIS eventually ended. The NTIS database was removed. And eventually, as Northern Light itself was bought out and then abandoned by Divine, Inc., USGOVSEARCH went the way of GovBot.

Within the government itself, several initiatives were underway to create a central, government-specific portal or search engine. GPO Access was one approach, but other agencies also tried developing their own. NTIS's FedWorld crossed agency boundaries. LC had its own directory. And then there was the WebGov initiative. Announced in September 1998, WebGov was supposed to be a central government-wide portal and was going to be up and running in thirty days. Due in part to fighting between agencies and in part to lack of funding, it never got off the ground (Brown, 2000).

It was not until two years later, in September 2000, that FirstGov rose out of the ashes of WebGov as a live, viable site. As O'Leary (2001) describes it in an early review, the initial version did have its problems and inaccuracies but, for the first time, the government had its own portal and search engine. Certainly, one major reason for this was that Eric Brewer of Inktomi had donated some of the technology.

FirstGov has become the most prominent government-sponsored, central access point for government information. And many other government sites link back to FirstGov. For a more detailed view of FirstGov, see Patricia Diamond Fletcher's article "Creating the Front Door to Government" in this issue of *Library Trends*.

Concerns and Challenges

One significant concern with the move away from multiple copies of print documents at many libraries to one electronic copy on an agency's Web site is that the loss of redundancy can easily lead to the permanent loss of the information content. If that one copy is removed, lost in a computer crash, or forgotten in a site upgrade, there is an increased potential that all future users will permanently lose access to it.

In the same vein, if an online copy is changed, due to necessary corrections or to political leadership changes, there may not be an archived

record of the original Web page. Without that, future historians, journalists, and others may not be able to identify what changes have been made and when they were made. Will a new administration in Washington try to purge Web pages from a previous administration? So far in the Internet age, we have only had the change from the Clinton administration to the Bush administration. And already there have been several efforts to remove older material.

Davis (2002) reports on an internal memo in the Department of Education that called not only for the department to remove outdated pages but also to remove items that might not reflect the current administration's political philosophy. While that process is still ongoing and the eventual fate of all the old pages is yet to be determined, this is exactly the kind of situation that is likely to become more frequent as the Web ages and as administrations change.

On another front, the Office of Management and Budget (OMB) proposed major changes to the requirement for government agencies to use GPO for printing publications (Procurement of Printing, 2002). Although the requirement to send publications to the Depository Library Program remains, great concerns have been raised about potential loss of many documents. Helfer (2003) argues that the fugitive documents problem will be made worse by this weakening of the Federal Depository Library Program and GPO. And the requirements in the OMB proposal that the Superintendent of Documents would bear the costs that have been legislatively mandated to be borne by the agencies themselves certainly could exacerbate the problem of fugitive and lost documents.

The nature of the Web makes it difficult for the government to keep many secrets because public information can be posted by almost anyone with a Web site. But if the information is never published or posted, that risk is averted. And even though most government agencies have moved toward putting many of the publications online, there are still other areas in which the government is reticent to publish certain documents or has actively removed them.

With the greater concern about terrorism since the September 11, 2001, attack on the World Trade Center, many government sites have actually removed information that was formerly available on the Internet. OMB Watch, another nongovernmental organization, has documented many such incidents on its Access to Government Information Post September 11th pages at http://www.ombwatch.org/article/archive/104/.

Yet even before September 11th, the military had expressed concern. The Deputy Secretary of Defense (1998) issued a memorandum calling for a department-wide review of information vulnerability on the Web. "All DoD components that establish publicly accessible Web sites are responsible for ensuring that the information published on those sites does not compromise national security or place DoD personnel at risk" (p. 1).

In another unusual situation completely separate from the Defense Department, the Department of the Interior was ordered to disconnect its systems from the Internet by a federal judge on December 5, 1999, due to vulnerabilities in Indian trust-fund databases. This meant that many sites such as the National Park Service, the Land Management and Reclamation Bureau, the Fish and Wildlife Service, the Minerals Management Service, and the Surface Mining Office all suddenly had no information or only very abbreviated information on their sites (Dizzard, 2002). The various agencies slowly were able to get approval from the court to bring their sites back up but, as Lisagor (2002) notes, 6 percent of the department's systems were still disconnected as of November 2002.

Unfortunately, no private organizations have mirrors of all the information that was on the vanished sites. Any information that was only accessible via an interactive database search is gone. But some of the more static pages have been available from services like the Internet Archive's Wayback Machine (http://www.archive.org) as long as a user knew the appropriate URL.

Conclusion

There is no doubt that the quantity and quality of government information on the Internet is a substantial resource for many kinds of users, from everyday U.S. citizens to advanced researchers of social trends. Government sites provide detailed data sets, satellite imagery, weather records and trends, tax forms, contractor opportunities, hazardous waste disposal pamphlets, elementary history lesson plans, consumer guides, proposed regulations, laws, court cases, speeches, testimonies, and so much more.

Compared to so much else on the Web that is of widely varying quality and often more concerned with selling something than providing accurate content, government sites offer a great wealth of information. Despite some of the concerns and challenges with documents not being published or even removed, there is still a vast quantity of government information freely available on the Web that is of great importance for scholars and researchers.

The desktop availability of the data in a wide variety of formats and from many different agencies is a boon for researchers who want access to data at their time of need rather than waiting for delivery of documents many days later. Finding the appropriate material can still be difficult, but search engines like FirstGov and directories and subject-oriented gateways are a great help.

Through the efforts of government agencies like the Library of Congress, the Government Printing Office, and many others, the Internet public is fortunate to have a substantial body of valuable information content available at the desktop for free, and at any time of the day from all over the world.

REFERENCES

Brown, D. (2000). Feds take another run at governmentwide portal. *Inter@ctive Week, 7,* 10–11.

Davis, M. R. (2002). No URL left behind? Web scrub raises concerns. *Education Week, 22,* 1, 26.

DeLoughry, T. J. (1994). Library of Congress announces plan to digitize several of its collections. *The Chronicle of Higher Education, 41,* A41.

Deputy Secretary of Defense. (1998). Information vulnerability and the World Wide Web [memorandum]. Washington, DC: Dept. of Defense. Retrieved April 14, 2003, from http://www.defenselink.mil/other_info/depsecweb.pdf.

Dizard, W. P. (2002). Interior struggles to restore its Web presence. *Government Computer News, 21,* 10.

Gordon-Murnane, L. (1999). The Federal Register free on GPO Access. *Searcher, 7*(6), 46.

GPO Access—Free, kind of. (1995). *Searcher, 3*(1), 32–33.

Hane, P. J. (1999). Northern Light Technology's USGOVSEARCH begins commercial operation with revised pricing plan. *Information Today, 16*(7), 22.

Helfer, D. S. (2003). The government battle over printing: OMB versus GPO and why it matters to libraries and the public. *Searcher, 11,* 53–55.

Kahle, B., Morris, H., Goldman, J., Erickson, T., & Curran, J. (1992). Interfaces for distributed systems of information servers. *Proceedings of the ASIS Mid-year Meeting,* 124–148.

Kaiser, J. (1998). PTO to let sun shine on patents. *Science, 281*(3), 7.

Library of Congress. (1995). New online public access to congressional information. [press release]. Retrieved April 14, 2003, from http://www.loc.gov/today/pr/1995/95-002.

Library of Congress goes online. (1993). *Link-Up, 10*(5), 22.

Lisagor, M. (2002). Still a disconnect at Interior. *Federal Computer Week.* Retrieved April 14, 2003, from http://www.fcw.com/fcw/articles/2002/1111/web-interior-11-11-02.asp.

Lynch, C. A., & Preston, C. M. (1990). Internet access to information resources. In M. E. Williams (ed.), *Annual review of information science and technology* (vol. 25, pp. 263–312). Amsterdam: Elsevier Science Publishers.

Minahan, T. (1994). GPO puts Congressional Record, Federal Register on the Internet. *Government Computer News, 13*(12), 8.

Notess, G. R. (1993). LC's debut on the Internet. *Database, 16*(5), 84–87.

Notess, G. R. (1994). SEC EDGAR and the Internet town hall. *Database, 17*(4), 77–81.

Notess, G. R. (2000). USGOVSEARCH: The federal Web, NTIS database, and more. *Online, 24,* 57–61.

O'Leary, M. (2001). FirstGov arrives on fast track. *Information Today, 18*(1), 17–19.

Procurement of printing and duplicating through the Government Printing Office (2002). *Federal Register, 67*(219), 68913–68918.

Rogers, M. (1991). Data Research offers network to all libraries. *Library Journal, 117*(19), 22.

Santo, B. (1995). Web proves patently useful. *Electronic Engineering Times, 869,* 126.

Creating the Front Door to Government: A Case Study of the *Firstgov* Portal

PATRICIA DIAMOND FLETCHER

ABSTRACT

FIRSTGOV IS THE U.S. FEDERAL PORTAL to government information and services. It was conceived by the Clinton administration in June of 2000 and launched in September 2000. A case study of the development of *Firstgov* indicated that top-level leadership, a small and committed project team, and the very condensed timeframe of the project were factors that contributed to the success of the portal. Another reason cited for the success of the *Firstgov* development was the U.S. federal information policy environment, a robust and evolving framework creating the climate for electronic government. An unusual feature of the project development was the donation of the Inktomi search engine for three years, an event that further enabled *Firstgov* to open its door on time and on budget. The portal continues today with funding and resources designed to ensure its future.

INTRODUCTION

The creation of the *FirstGov* federal Web portal represents a dramatic new way of doing business for government. The portal itself represents a major change in how the government will interact with its customers—citizens, businesses, and other governments. The longstanding oxymoron, "technical innovation in government," has been challenged with the development of *FirstGov*. This application, created during the Clinton administration, has paved the way for the e-government strategy of the Bush administration. The goal is for *FirstGov* to serve as the gateway to all U.S. government information. It provides the most comprehensive search of

Patricia Diamond Fletcher, Associate Professor, Policy Sciences Graduate Program, University of Maryland Baltimore County, Baltimore, MD 21250
LIBRARY TRENDS, Vol. 52, No. 2, Fall 2003, pp. 268–281

government documents and services anywhere on the Internet. The creation of this portal was informed by policy designed to create an electronic government. It was forged by a unique partnership between the public and private sectors, which enabled it to be up and running in ninety days—a major feat for government. The story of this development was captured in a case study funded by the National Science Foundation under the Digital Government program.

The U.S. Federal Government Information Environment

Policy is a critical tool for framing the operational environment for government (Dawes et al., 1999; Fletcher & Westerback, 1999). Policy related to information and the management of information resources has had a defining influence on the evolution from a paper-based, to a computer-based, to an electronic government in the United States. When viewed from the perspective that the U.S. federal government is the world's largest creator, disseminator, and user of information, the criticality of having a strong policy framework is obvious. Harlan Cleveland (1986) asserted that "government is information." The importance and value of information to government mandates a high level of attention to ensure that it will be utilized for the public good. This policy framework serves to highlight and unify information issues such as management, planning, privacy, security, access, property rights, and electronic commerce.

Policy Creating an Electronic Government

The Government Paperwork Elimination Act (GPEA) (P.L. 105–277), signed into law October 21, 1998, represented the Clinton administration's intent to move quickly to a federal government that offered comprehensive electronic access and services. GPEA was a major legislative endorsement of electronic government. It required the federal executive agencies, no later than October 21, 2003, to allow individuals and businesses that interact with federal agencies the opportunity to do so electronically. GPEA more importantly mandated that electronic records and their electronic signatures were to have the full force of legal effect and validity. It encouraged federal agencies to promote an electronic information-management environment more akin to electronic commerce models, including electronic transactions, recordkeeping, filing, maintenance, submission, and archiving. This opened up a wide array of possible types of electronic information interactions between government and the public. The submission of bids and proposals for government contracts; applications for licenses, loans, and benefits; requests for government records; receipt of benefits such as social security; online procurement; and citizen interaction in legislation are but a few examples of the new applications for which GPEA created the policy environment.

The high-level management policy environment for electronic government is set forth in S.803, the E-Government Act of 2001, introduced by Senator Joseph Lieberman (D-CT). While it was not successful in 2001, an amended version of the act was reintroduced and reported out of committee on March 21, 2002. With strong congressional support, this bill was passed by the Senate on November 15, 2002, mere hours after the House had approved the measure. The amended version of the E-Government Act (P.L. 107–30) sets up a broad policy framework for an electronic government strategy that will enable citizens to access their government information and services electronically, over the Internet. The act recognizes the effect the Internet has already had on U.S. society and seeks to avail both government and citizens of the benefits already being realized by businesses and individual Internet users. The act further includes the creation of a federal chief information officer (CIO) housed in the Executive Office of Management and Budget (OMB) and the establishment of an Office of Electronic Government housed in OMB. The federal chief information officer is to be appointed by the president with the advice and consent of the Senate. The creation of an Office of Electronic Government is to ensure that electronic initiatives are sound investments and, more importantly, that these new e-government initiatives are cross-agency in nature. This is a serious effort to dismantle the unwieldy "stovepipe" structure that is predominant today across government. Cross-agency initiatives are seen as reducing the information burden on the public, while making access simplified, universal, and not time limited.

A critical aspect of the act is funding. It appropriates $45 million for funding of electronic government initiatives in the current fiscal year (Executive Office of the President, 2003). In subsequent years, the Office of Electronic Government will have a total of $345 million to be expended over five years. This is a needed shot-in-the-arm for electronic government development, which had been appropriated a mere $5 million for fiscal year 2002. Some of the funds will go to improvements on the *FirstGov* portal. The development of a subject-based taxonomy for users is a vital component of the changes envisioned for *FirstGov.* This will move the portal, and the federal government, away from the current agency-based locus of information.

There are many other laws that frame the electronic government environment. The development of an information resources management environment has been a slow and deliberate process in federal government, and it created the framework for an electronic government to flourish. The Commission on Federal Paperwork, created under the Ford administration, was the bellwether for the development of many of the following laws related to the electronic management of information. Some of the key laws that have enabled an electronic government to evolve are the following:

- The Paperwork Reduction Act of 1995 (44 U.S.C. chapter 35);
- The Clinger-Cohen Act of 1996 (40 U.S.C. 1401(3));
- The Government Information Security Reform Act of 2000 (P.L. 106–398):
- The Computer Security Act of 1987, as amended (P.L. 100–235, 15 U.S.C.);
- The Privacy Act of 1974 (5 U.S.C. 552a);
- The Computer Matching and Privacy Protection Act of 1988 (P.L. 100–503); and
- The Telecommunications Act of 1996 (P.L. 104–104).

Another defining policy statement in support of electronic government was set out in the *President's Management Agenda* of 2001. The Bush administration developed five government-wide goals for its tenure, one of which is the expansion of an electronic federal government. Thus, the imprimatur for continuing evolution of an electronic government was set. A high-level task force was created from an July 18, 2001, memo (OMB-01–28) which called for the task force, informally named "quicksilver," to develop the priority strategic actions needed to enable electronic government. The group was in service by August of 2001, and it quickly set out an ambitious agenda for electronic government. This agenda was reported to the President's Management Council on October 3, 2001—truly Internet speed! The initial electronic government agenda was further refined and formalized in the February 27, 2002, release of the *E-Government Strategy* (Executive Office of the President, 2002). This strategy created a vision that is citizen centered, results oriented, and market based in nature. It mandates cross-agency sharing of data to simplify access to government and to reduce information resources expenses across government agencies. The strategy focuses on four groups of end-users—government-to-citizen, government-to-business, government-to-government, and intragovernmental—to improve internal efficiency and accountability of federal agencies. An initial thirty-four projects were singled out for the first round of funding, with completion dates scheduled no later than eighteen to twenty-four months. All approved projects represented cross-agency applications. The haste to get them online is a further measure of the importance of electronic government to the administration's overall policy.

Policy Creating the Firstgov Portal

The use of the *FirstGov* portal as an anchor for these more agency- or service-based applications is a key component of the electronic government strategy outlined above. The portal both complements and enables the information policy framework of the federal government. *Firstgov* is seen today as a key player in the continued management and development of the e-government initiatives. The Clinton administration's strong support

of the use of information technology in government set the stage for the eventual adoption of a portal model of Internet use. Clinton, in the instantiation of the National Performance Review (later renamed the National Partnership for Reinventing Government[1]) created an Internet-enabled environment for government early on in his first term of office. Through the adoption of increasingly sophisticated information technology, the federal government was poised to utilize the Internet in its daily practice.

The December 17, 1999, presidential memo on "Electronic Government" was the first policy-level indication that the then Clinton administration wanted to create a "one-stop" access point for government information and services. This commitment was reaffirmed in the first presidential Internet address on June 24, 2000. In his address, President Clinton stated that a government portal for information and services was to be open for business within ninety days of the Internet address, thus giving the policy not only "teeth" but also a major challenge. It was clear that Clinton saw *FirstGov* as a legacy he wanted to leave the American public when he stepped down from office in January of 2001. The above detailed laws—coupled with the strong presidential support and direction—created a policy environment that was predisposed to the successful development of a federal information and service portal.

The Case Study

A case study approach to understanding the *FirstGov* implementation was seen as providing the richest data. This project was the first of its kind in the federal government and spanned both public and private sectors in a new model of partnership. By approaching *FirstGov* as a case, we were able to investigate six dimensions of a preliminary conceptual model of electronic government collaborative developments. The dimensions included in the model are:

- Political, social, economic, and cultural environment;
- Institutional, services sector, and technological environment;
- Characteristics and objectives of public and private partners;
- The collaboration process;
- The collaboration methods; and
- Performance.[2]

The case study interviews were conducted by a team of researchers at the University of Maryland, Baltimore County, in the summer of 2001. We interviewed the key participants in the development of the portal—both public and private sector partners and stakeholders. The enabling policies were analyzed, along with any relevant documentation on the project complied by the *FirstGov* team. The testimonies from the House Subcommittee on Government Management, Information, and Technology hearing on

FirstGov (October 2, 2000) were part of the documentation analyzed for the case. The development process for *FirstGov* received considerable attention by the federal information community press as well, and relevant articles from magazines such as *Federal Computer Week* and *Government Computer News* were scanned on a regular basis for stories about the project. Data were coded based on a scheme developed and pretested by the research team at the Center for Technology in Government at SUNY Albany.

The Creation and Implementation of *FirstGov*

FirstGov was launched September 22, 2000, with an initial size of 47 million U.S. federal government Web pages. *FirstGov*, the only official U.S. government Web portal, is described as a single, trusted point-of-service for U.S. citizens and businesses to gain entry to federal services and information resources. The initial vision for *FirstGov* was to be a high-speed, twenty-four hours a day, seven days a week, user-friendly entry point to every online resource, be it information, data, or service, offered by the federal government of the United States. *Firstgov* was also envisioned as the vehicle to reduce government bureaucracy substantively, create a more responsive and customer-focused government, and enable new and more active citizen participation in democratic processes.

FirstGov serves as an example of a unique public-private partnership to provide electronic government services and information to the public. This project represented an entirely new venture for the U.S. federal government. It was created to cut across agency and departmental stovepipes and to centralize the location for retrieval of government information and services, with government agencies traditionally being averse to either activity. While a number of portal-type applications were developed under the National Performance Review (e.g., http://www.students.gov, http://www.seniors.gov, and http://www.workers.gov), *FirstGov* represented a project on a much larger scale, with its scope being the entire federal government.

To provide ongoing direction to the project, the President's Management Council (PMC) established a FirstGov.gov Board of Directors, which consisted of eight members from the PMC and three members of the Federal Chief Information Officers (CIO) Council. The board was charged with responsibility for coordinating project issues across the executive, legislative, and judicial branches of government. The daily development and management of the portal were turned over to the U.S. General Services Administration (GSA), which staffed a *FirstGov* project team to lead the effort. This team, in turn, managed a $4 million, two-year contract to create, operate, and maintain the Web site. The GSA was a key partner in the development process. It provided the wherewithal, the organizational resources, and a good number of the people to work on *FirstGov*. The *FirstGov* team was created as a collateral model of the

organization, one that used the resources of the larger agency but worked outside standard operating procedures as needed. Thus, team expertise and enthusiasm were not hampered by the red tape of bureaucracy. The then CIO at the GSA was credited with being a driving force behind the project's success. He was referred to as an advocate, a proselytizer, and a very visible champion for *FirstGov* throughout its development and implementation.

The above-mentioned contract did not cover services such as redesigning the Web site or changing its hosted location. It also did not cover the development or use of an electronic search function—a critical aspect of this project. That search function was offered free of charge for an initial three-year period by the Federal Search Foundation (Fed-Search). Fed-Search was the nonprofit corporation developed by Dr. Eric Brewer, cofounder and chief scientist for the Inktomi Corporation, to channel the donated search engine to *FirstGov*. In setting up this corporation, Dr. Brewer also envisioned that it would attract other private sector partners who would be eager to donate some technology component or service to this innovative and potentially profitable project. A memorandum of understanding with the GSA, on behalf of the PMC and the FirstGov.gov Board, and Dr. Brewer cemented this generous donation of a world-class search engine. It was believed by members of the project team that this donation was one of the key critical elements that enabled the project to be completed on time. It is interesting to note here that this same donation was the cause of considerable angst in the software industry, which feared that, when the three-year donation period was over, Inktomi would have an unfair competitive advantage over other potential vendors vying for the contract.

The Federal CIO Council was also a partner in the project. It was used as a source of knowledge and expertise on government agencies and information technology. The agency CIOs were also coopted to be change agents to convince agency personnel of the necessity of being a part of *FirstGov* and not a protagonist. Thus, the CIOs were able to provide support for the cross-agency approach to information presentation and dissemination—a vital characteristic of the *FirstGov* portal. The Federal CIO Council also assisted the project by providing some funding for the first-year development and maintenance of the portal. They literally passed the hat among twenty-two federal agencies to keep the project alive.

Everyone involved in the development and implementation of *FirstGov* expressed a sense of dedication to and belief in what they were doing. The sense of importance, high-level commitment, and urgency was transmitted through all the partners, who pulled together to make the project a success. This was not a typical government project, mired in procurement and acquisition regulations and constrained by the federal budget, although it was noted repeatedly that the small initial budget was a hin-

drance to the development team. The *FirstGov* project was much more like that of a start-up "dot-com" fueled by the energy and engagement of its members and their belief in the project's goals and objectives.

Another important motivator for the partners was that *FirstGov* was seen as a necessary and important public service. The strong information policies of the federal government focused on information creation, dissemination, and records management and archiving. The development of a government-wide portal was but one step in the move to an electronic government—a government that would facilitate the access and dissemination of information.

Critical Success Factors for the Project

Leadership was from the very top, the president of the United States. Clinton was a champion for using information technology to enable better, smarter, faster government services and information dissemination. The top-level attention from the Executive Office of the President was one of the critical success factors that enabled the portal to be "open for business" in such an unprecedented amount of time—ninety days. The criticality of such top-level support has long been addressed in the research literature (Kraemer & King, 1977; Fletcher et al., 1992; Norris & Kraemer, 1994; Norris & Kraemer, 1996; Fletcher, Holden, & Norris, 2001). The pervasive impact of this variable and its effect on the success of such a monumental information technology project was well demonstrated by the *FirstGov* project.

The management of the project, in the hands of the U.S. General Services Administration, was a facilitating factor in the project's perceived success. The GSA team members were tirelessly dedicated to the project because "they knew it was right." And many saw the small size of the team as a success factor. The size enabled it to be fast and flexible. All of the people interviewed credited the following as well to the successful launch of *FirstGov:*

- The president's memo of December 17, 1999, on "Electronic Government";
- The passage of the Government Paperwork Elimination Act in 1998;
- The donation of the Inktomi search engine for a three-year period;
- The small size of the project team; and
- The compressed time frame—ninety days—in which to develop and implement *FirstGov.*

These factors created the necessary top-level support, the policy framework, and the sense of commitment and urgency to have a successful project. A general theme heard echoed among the respondents was that *FirstGov* was successful because of personality, commitment, and a good

team. While many of the noted critical success factors come as no surprise, the fact that the very brief development schedule was seen as positive represented something new for the federal government. Unlike most information technology projects in government, where procurement and acquisition law often contribute to lengthy, drawn-out, and costly information technology developments, *FirstGov* was not subject to many of these instances of red tape. The requirement of a ninety-day project development meant that, to be successful, the team had to creatively, while legally, procure the necessary technology to launch the portal on time. This created a sense of urgency that spurred the team to exceed their performance expectations.

The critical success factors sum up the components of the partnership and the development activities well. There was a policy environment in place that was conducive to creating an electronic government portal. There was presidential support and a committed project team. The donation of a search engine significantly cut down the time and expense needed to assess and procure or create a search engine with the necessary capabilities for the portal. This was a very visible, high-impact project, and there was considerable scrutiny from stakeholders and from the press. These pressures served to motivate the team to work harder and faster than many anticipated. *Firstgov* was launched on time and on budget to visible fanfare.

Assessment of the Portal

The *FirstGov* initiative was seen by many as transformational to the conduct of government. It has received numerous awards since the portal went live in 2000. It has also been embraced by the Bush administration, with Vice President Cheney launching the redesigned portal in February of 2002. Among the awards it has been given are:

- *Yahoo! Internet Life* magazine's Fifty Most Incredibly Useful Sites, July 2002;
- Pioneer Award, E-Gov 2002, June 2002, and April 2001;
- Industry Advisory Council, E-Gov, and the Federal Chief Information Officer Council's Excellence.Gov Award Finalist, January 2002;
- *Government Executive* magazine's 2001 Grace Hopper Government Technology Leadership Award, December 2001;
- 2001 Innovations in American Government Award Finalist, August 2001 and Semifinalist, April 2001;
- Federation of Government Information Processing Council's Intergovernmental Solutions Award, June 2001;
- 2001–2002 Golden Web Award, May 2001;
- Azimuth Award for supporting federal information technology went to Dave Barram, former GSA administrator, and Eric Brewer, for their part in *FirstGov*.gov, March 2001;

- FOSE and Chief Information Officers Council of Excellence Award, March 2001;
- Vice President's Hammer Award for Reinventing Government, January 2001. (Awards and recognition of *FirstGov,* n.d.)

Today (December 2002), there are more than 51 million Web pages at *FirstGov* from more than 2,000 Web sites, not only from the federal government but also from the District of Columbia, state governments, and the U.S. territories. Pages accessible on *FirstGov* are, by-and-large, not available on other commercial Web sites. The redesigned Web site is arranged by three gateways: citizen, business, and government. It is informational and transactional, enabling users to conduct business with government via the Internet. Transactions are available for citizen-to-government, business-to-government, and government-to-government processes. You can find and apply for government jobs, electronically pay an employee's child support obligation, electronically file for patent and trademarks, purchase government supplies, apply for federally guaranteed student loans, buy stamps, change your address, and a whole host of other activities that used to require bricks-and-mortar, paper-and-pencils. This is the twenty-four–hour access and convenience that was the goal of *FirstGov* when it went online.

In a study conducted by Stowers (2002), the author noted that the design and content of the site were both well thought-out and effective for the end user. Stowers described *FirstGov* as "strongly citizen focused" and gave high marks to its portfolio-type user gateways. The portal meets one of the most important criteria that users ask for in a government Web site—the ability to communicate with elected officials (Matthews, 2002), which is in line with Stowers's assessment above that *FirstGov* has a strong citizen orientation.

Firstgov has done some of its own soul-searching as well. In a survey administered to gauge customer satisfaction (May 2002) first-time users of the portal indicated that they were much more likely to revisit *FirstGov* than they had been prior to its February 2002 redesign. This was the most significant finding of the survey. Return users to the portal noted that it was easier to find information and that they more often now recommended *FirstGov* to others as a search engine.

Of course, as with anything done by the government, not all reviews of *FirstGov* have been favorable. The portal has been criticized as not accessible to end-users, little more than a table of contents to government, not meeting many project deadlines and, most recently and visibly, it has received much adverse publicity for awarding the new search engine contract to a Norwegian company. This award was greeted with dismay and outright antagonism, as many felt the search engine for the premier U.S. government Web site should be a U.S. company. However, *FirstGov* has

gone ahead with this award and Fast Search and Transfer will provide the search services for the next five years at a projected cost of $1.85 million a year (*Federal Computer Week,* 2002). The selection of the Oslo-based company did, however, dispel the fear of many in the software industry that Inktomi, with its initial donation of the *FirstGov* search engine, had an unfair competitive advantage. While Inktomi bid for the new procurement, it was not chosen.

Probably one of the most cogent comments that can be made about *FirstGov* at this time is that it is a work in progress, as are all government Web sites today. With the completion of the $350,000 site redesign, and the $85,000 contract to UserWorks to test the usability of the site extensively, *FirstGov* appears to be ready to learn from its past. The newly reorganized operating structure for the *FirstGov* staff is another indication that the administration is supporting major changes in operating procedures to better offer information and service access through this portal. The General Services Administration has reorganized the *FirstGov* office into a consolidated customer-focused unit—the Office of Citizen Services and Communications. *Firstgov* is an integral part of this new office, enabling the GSA to act as a front door to the services and information sought by U.S. citizens. In support of this focus, the GSA has designated their e-government activities as one of their three 2003 budget themes, thus providing the needed resources. The president's e-government strategy, with its recent funding of twenty-four new cross-agency initiatives, also lends considerable support to the future of *FirstGov.* The portal is to be a major player in the development and implementation of the e-government strategy. It has also been awarded a portion of OMB's innovative e-government projects fund (*Federal Computer Week,* 2002), with a focus on e-authentication and content management of the portal. *Firstgov* will also receive a significant portion of the fiscal year 2003 information technology budget, set at $52 billion.

The recently enacted E-government Act of 2002 also creates a rosy future for *FirstGov.* The act sets aside a fund of $345 million to be administered by the GSA over the next four years in support of e-government projects. As noted above, the oversight of *FirstGov* is in the GSA, a fortuitous location for the e-government portal. Thus, the future for this portal is bright. The top-level support for electronic government has carried over from the Clinton to the Bush administration. The policy environment supports its continued development and maintenance. The American public is online and taking advantage of government Web sites. A recent report from the Council for Excellence in Government (Hart-Teeter, 2002) indicated that 76 percent of all Internet users and 51 percent of Americans have accessed a government Web site. It also noted that, overall, Americans are more positive in their outlook toward electronic government than they were in the previous year, and that they had high expectations for government as it went online. Government Web sites that duplicate the ease

and usability of the "dot-coms" are expected, and *FirstGov,* with its redesign and its responses to user surveys, is well aware of this expectation. Further, *FirstGov* has won numerous awards over the past three years and has strong visibility and usage. It is poised to play a critical role in both the implementation of the *President's Management Agenda* and the electronic government initiatives funded in the 2003 budget of the United States. The 2003 budget recognizes that the U.S. government will mix its use of Internet and physical assets to become a "click and mortar" enterprise. The agencies that serve citizens, businesses, internal federal government functions, and intergovernmental needs will thus become more accessible, effective, and efficient. In adopting a "click and mortar" model, the federal government will use the best practices of industry. The Bush administration's goal is that services and information sought by citizens will rarely be more than three clicks away from end users.

Conclusion

A final thought here has to do with the imperative of access to government information. This principle has been the drive behind information policy and management in federal government. But it is hindered by the perpetual inefficiencies of data redundancy, data duplication, and data error that abound in government information systems. The creation of *FirstGov* does not remediate these age-old problems with access to data. It does not mean that all government information will reside in one format, in one location. Rather, *FirstGov* makes use of existing federal agency databases for its content. It is no secret that these agency Web sites are often less than optimal (McClure, Sprehe, & Eschenfelder, 2001). Federal agency Web site development began with the agencies putting their paper products online and is only now slowly moving toward a reengineering orientation for the online environment. Thus, in many instances, we are receiving the electronic version of our paper government rather than seeing government reengineered for an electronic environment and citizenry.

There are further complications and complexities when we add into this mix the state and local government Web sites. All U.S. state governments have Web sites, many of these being all-inclusive gateways to state government. One need only go to North Star, the official home of Minnesota government (http://www.state.mn.us/) or AccessWashington (http://access.wa.gov/) to see innovative and diverse approaches to online information access and service delivery. Cities such as New York and Chicago are also making use of the portal concept, offering a "mygov.gov" approach for their users. In respect to the diffusion curve, the state and local governments appear to be in the lead, and *FirstGov* can take some lessons learned and best practices from these innovative and citizen-centric applications.

An additional complexity in creating an all-inclusive U.S. government portal is that state and local governments operate under different

information-management policies and environments when it comes to public records, privacy, security, and infrastructure concerns. There are many important questions to be thought through and problems to be resolved as we move forward in our electronic world. Access and usability need to be kept in the forefront of development goals—maybe not always compatible with state and local needs, but essential to the success of *FirstGov.* Our portal to electronic government has been constructed—what remains to be seen is how it will develop into our front door to government.

NOTES

1. For a more robust description and assessment of the information policy environment that framed the National Performance Review, see Fletcher & Westerback (1999).
2. A detailed explanation of the model and the major research results can be found at http://www.cefrio.qc.ca/english/activites_symp.cfm, from an International Conference on Public-Private Partnerships for Improved Government Performance, October 24–25, 2002, Quebec City, Canada.

REFERENCES

Awards and recognition of FirstGov. (n.d.). Retrieved January 2, 2003, from http://www.firstgov.gov/About/Awards.shtml.

Cleveland, H. (1986). Government is information (But not vice versa). *Public Administration Review, 46,* 605–607.

Dawes, S. S., Bloniarz, P. A., Kelly, K. L., & Fletcher, P. D. (1999). *Some assembly required: Building a digital government for the 21st century.* Albany, NY: Center for Technology in Government.

Executive Office of the President, Office of Management and Budget. (2002). *E-government strategy: Simplified delivery of services to citizens.* Retrieved March 1, 2002, from http://www.whitehouse.gov/omb/inforeg/egovstrategy.pdf.

Executive Office of the President, Office of Management and Budget. (2001). *The president's management agenda.* Retrieved March 15, 2002, from http://www.whitehouse.gov/omb/budintegration/pma_index.html.

Executive Office of the President, Office of Management and Budget. (n.d.). *Budget of the United States government, fiscal year 2003.* Retrieved June 6, 2002, from http://www.whitehouse.gov/omb/budget/index.html.

Federal Computer Week. (2002, April 22). Things you need to know about the e-government act. Retrieved January 2, 2003, from http://www.fcw.com/fcw/articles/2002/1202/Egovactchart.pdf.

Federal Computer Week. (2002, March 11). FirstGov search returns surprise result. Retrieved January 2, 2003, from http://www.fcw.com/fcw/articles/2002/0311/news-search-03-11-02.asp.

Firstgov.gov homepage. (n.d.). Retrieved December 12, 2002, from http://www.firstgov.gov.

Fletcher, P. D. (2003). Portals and policies: Implications of electronic access to U.S. federal government information and services. In G. David Garson (Ed.), *Digital government: Principles and best practices.* Hershey, PA: Idea Group Press.

Fletcher, P. T., Bretschneider, S. I., & Marchand, D. A. (1992). *Managing information technology: Transforming county governments in the 1990s.* Syracuse, NY: Syracuse University, School of Information Studies.

Fletcher, P. D., & Westerback, L. (1999). Federal information policy: Management to measurement. *Journal of the American Society for Information Science,* Special Issue on the National Information Infrastructure, *50*(4), 299–304.

Hart-Teeter for The Council for Excellence in Government. (2002). *E-government to protect, connect, and serve us.* Retrieved March 20, 2002, from http://excelgov.xigroup.com/displayContent.asp?Keyword=ppp022602.

Hasson, J. (2002). E-gov agenda takes shape. *Federal Computer Week.* Retrieved January 2, 2003, from http://www.fcw.com/fcw/articles/2002/1202/news-egov-12-02-02.asp.

Kraemer, K. L., & King, J. L. (1977). *Computers and local government.* New York: Praeger.

Matthews, W. (2002). FirstGov search returns surprise result. *Federal Computer Week.* Retrieved December 20, 2002, from http://www.fcw.com/fcw/articles/2002/0311/news-search-03-11-02.asp.

Matthews, W. (2002). GSA debuts friendlier FirstGov. *Federal Computer Week.* Retrieved January 2, 2003, from http://www.fcw.com/fcw/articles/2002/0304/news-first-03-04-02.asp.

McClure, C. R., Sprehe, T., & Eschenfelder, K. (2001). *Performance measures for federal agencies: Final report.* Retrieved December 13, 2001, from http://www.access.gpo.gov/su_docs/index.html.

Norris, D. F., & Kraemer, K. L. (1994). Leading edge computer use in U.S. municipalities. *ICMA Special Data Issue* (pp. 1–65). Washington, D.C.: International City Management Association.

Norris D. F., & Kraemer, K. L. (1996). Mainframe and PC computing in American cities: Myths and realities. *Public Administration Review, 56,* 568–576.

Presidential Memo of December 17, 1999. *Electronic government.* Retrieved January 3, 2000, from http://whitehouse.gov.

Remarks by the president in the first Internet Webcast. (2000). Retrieved October 10, 2000, from http://www.whitehouse.gov/WH/new/html/internet2000-02-24text.html.

Stowers, G. N. L. (2002). The state of federal Websites: The pursuit of excellence. The Price-waterhouse Coopers Endowment for the Business of Government. Retrieved December 1, 2002, from http://www.endowment.pwcglobal.com/pdfs/StowersReport0802.pdf.

U.S. General Services Administration. Office of Citizen Services and Communications. (n.d.). Retrieved January 2, 2003, from http://www.gsa.gov/Portal/content/orgs_content.jsp?channelId=-13536&contentOID=22916&contentType=1005.

The Invisible Web: Uncovering Sources Search Engines Can't See

CHRIS SHERMAN AND GARY PRICE

ABSTRACT

THE PARADOX OF THE INVISIBLE WEB is that it's easy to understand why it exists, but it's very hard to actually define in concrete, specific terms. In a nutshell, the Invisible Web consists of content that's been excluded from general-purpose search engines and Web directories such as Lycos and LookSmart—and yes, even Google. There's nothing inherently "invisible" about this content. But since this content is not easily located with the information-seeking tools used by most Web users, it's effectively invisible because it's so difficult to find unless you know exactly where to look.

In this paper, we define the Invisible Web and delve into the reasons search engines can't "see" its content. We also discuss the four different "types" of invisibility, ranging from the "opaque" Web which is relatively accessible to the searcher, to the truly invisible Web, which requires specialized finding aids to access effectively.

The visible Web is easy to define. It's made up of HTML Web pages that the search engines have chosen to include in their indices. It's no more complicated than that. The Invisible Web is much harder to define and classify for several reasons.

First, many Invisible Web sites are made up of straightforward Web pages that search engines could easily crawl and add to their indices but do not, simply because the engines have decided against including them. This is a crucial point—much of the Invisible Web is hidden *because search engines*

Chris Sherman, President, Searchwise, 898 Rockway Place, Boulder, CO 80303; Gary Price, Librarian, Gary Price Library Research and Internet Consulting, 107 Kinsman View Circle, Silver Spring, MD 20901.

LIBRARY TRENDS, Vol. 52, No. 2, Fall 2003, pp. 282–298

Partially excerpted from *The Invisible Web: Uncovering Information Sources Search Engines Can't See* by Chris Sherman and Gary Price (CyberAge Books, 0–910965-51-X).

have deliberately chosen to exclude some types of Web content. We're not talking about unsavory "adult" sites or blatant spam sites—quite the contrary! Many Invisible Web sites are first-rate content sources. These exceptional resources simply cannot be found using general-purpose search engines because they have been effectively locked out.

There are a number of reasons for these exclusionary policies, many of which we'll discuss. But keep in mind that, should the engines change their policies in the future, sites that today are part of the Invisible Web will suddenly join the mainstream as part of the visible Web. In fact, since the publication of our book *The Invisible Web: Uncovering Information Sources Search Engines Can't See* (Medford, NJ: CyberAge Books, 2001, 0–910965-51-X/softbound), most major search engines are now including content that was previously hidden—we'll discuss these developments below.

Second, it's relatively easy to classify some sites as either visible or invisible based on the technology they employ. Some sites using database technology, for example, are genuinely difficult for current generation search engines to access and index. These are "true" Invisible Web sites. Other sites, however, use a variety of media and file types, some of which are easily indexed and others that are incomprehensible to search engine crawlers. Web sites that use a mixture of these media and file types aren't easily classified as either visible or invisible. Rather, they make up what we call the "opaque" Web.

Finally, search engines could theoretically index some parts of the Invisible Web, but doing so would simply be impractical, either from a cost standpoint, or because data on some sites is ephemeral and not worthy of indexing—for example, current weather information, moment-by-moment stock quotes, airline flight arrival times, and so on. However, it's important to note that, even if all Web engines "crawled" everything, an unintended consequence could be that, with the vast increase in information to process, finding the right "needle" in a larger "haystack" might become more difficult. Invisible Web tools offer limiting features for a specific data set, potentially increasing precision. General engines don't have these options. So the database will increase but precision could suffer.

Invisible Web Defined

The Invisible Web: Text pages, files, or other often high-quality authoritative information available via the World Wide Web that general-purpose search engines cannot, due to technical limitations, or will not, due to deliberate choice, add to their indices of Web pages. Sometimes also referred to as the "deep Web" or "dark matter."

This definition is deliberately very general, because the general-purpose search engines are constantly adding features and improvements to their services. What may be invisible today may suddenly become visible

tomorrow, should the engines decide to add the capability to index things that they cannot or will not currently index.

Let's examine the two parts of this definition in more detail. First, we'll look at the technical reasons search engines can't index certain types of material on the Web. Then we'll talk about some of the other nontechnical but very important factors that influence the policies that guide search engine operations.

At their most basic level, search engines are designed to index Web pages. Search engines use programs called crawlers (a.k.a., "spiders" and "robots") to find and retrieve Web pages stored on servers all over the world. From a Web server's standpoint, it doesn't make any difference if a request for a page comes from a person using a Web browser or from an automated search engine crawler. In either case, the server returns the desired Web page to the computer that requested it.

A key difference between a person using a browser and a search engine spider is that the person can manually type a URL into the browser window and retrieve the page the URL points to. Search engine crawlers lack this capability. Instead, they're forced to rely on links they find on Web pages to find other pages. If a Web page has no links pointing to it from any other page on the Web, a search engine crawler can't find it. These "disconnected" pages are the most basic part of the Invisible Web. There's nothing *preventing* a search engine from crawling and indexing disconnected pages—but without links pointing to the pages, there's simply no way for a crawler to discover and fetch them.

Disconnected pages can easily leave the realm of the invisible and join the visible Web in one of two ways. First, if a connected Web page links to the disconnected page, a crawler can discover the link and spider the page. Second, the page author can request that the page be crawled by submitting it to "search engine add URL" forms.

Technical problems begin to come into play when a search engine crawler encounters an object or file type that's not a simple text document. Search engines are designed to index text and are highly optimized to perform search and retrieval operations on text. But they don't do very well with nontextual data, at least in the current generation of tools.

Some engines, like AltaVista and Google, can do limited searching for certain kinds of nontext files, including images, audio, or video files. But the way they process requests for this type of material are reminiscent of early Archie searches, typically limited to a filename or the minimal alternative (ALT) text that's sometimes used by page authors in the HTML image tag. Text surrounding an image, sound, or video file can give additional clues about what the file contains. But keyword searching with images and sounds is a far cry from simply telling the search engine to "find me a picture that looks like Picasso's 'Guernica'" or "let me hum a few bars

of this song and you tell me what it is." Pages that consist primarily of images, audio, or video, with little or no text, make up another type of Invisible Web content. While the pages may actually be included in a search engine index, they provide few textual clues as to their content, making it highly unlikely they will ever garner high relevance scores.

Researchers are working to overcome these limitations. Google, for example, has experimented with optical character recognition processes for extracting text from photographs and graphic images, in its experimental Google Catalogs project (*Google Catalogs,* n.d.). While not particularly useful to serious searchers, Google Catalogs illustrates one possibility for enhancing the capability of crawlers to find Invisible Web content.

Another company, Singingfish (owned by Thompson) indexes audio streaming media and makes use of metadata embedded in the files to enhance the search experience (*Singingfish,* n.d.). ShadowTV performs near real-time indexing of television audio and video, converting spoken audio to text to make it searchable (*Shadow TV,* n.d.).

While search engines have limited capabilities to index pages that are primarily made up of images, audio, and video, they have serious problems with other types of nontext material. Most of the major general-purpose search engines simply cannot handle certain types of formats. When our book was first written, PDF and Microsoft Office format documents were among those not indexed by search engines. Google pioneered the indexing of PDF and Office documents, and this type of search capability is widely available today.

However, a number of other file formats are still largely ignored by search engines. These formats include:

- Postscript,
- Flash,
- Shockwave,
- Executables (programs), and
- Compressed files (.zip, .tar, etc.).

The problem with indexing these files is that they aren't made up of HTML text. Technically, most of the formats in the list above can be indexed. AlltheWeb.com, for example, recently began indexing the text portions of Flash files, and Google can follow links embedded within Flash files.

The primary reason search engines choose not to index certain file types is a business judgment. For one thing, there's much less user demand for these types of files than for HTML text files. These formats are also "harder" to index, requiring more computing resources. For example, a single PDF file might consist of hundreds or even thousands of pages, so even those engines that do index PDF files typically ignore parts of a document

that exceed 100K bytes or so. Indexing non-HTML text file formats tends to be costly. In other words, the major Web engines are not in business to meet every need of information professionals and researchers.

Pages consisting largely of these "difficult" file types currently make up a relatively small part of the Invisible Web. However, we're seeing a rapid expansion in the use of many of these file types, particularly for some kinds of high-quality, authoritative information. For example, to comply with federal paperwork reduction legislation, many U.S. government agencies are moving to put all of their official documents on the Web in PDF format. Most scholarly papers are posted to the Web in Postscript or compressed Postscript format. For the searcher, Invisible Web content made up of these file types poses a serious problem. We discuss a partial solution to this problem later in this article.

The biggest technical hurdle search engines face lies in accessing information stored in databases. This is a huge problem, because there are thousands—perhaps millions—of databases containing high-quality information that are accessible via the Web. Web content creators favor databases because they offer flexible, easily maintained development environments. And increasingly, content-rich databases from universities, libraries, associations, businesses, and government agencies are being made available online, using Web interfaces as front-ends to what were once closed, proprietary information systems.

Databases pose a problem for search engines because every database is unique in both the design of its data structures and its search and retrieval tools and capabilities. Unlike simple HTML files, which search engine crawlers can simply fetch and index, content stored in databases is trickier to access, for a number of reasons that we'll describe in detail below.

Search engine crawlers generally have no difficulty finding the interface or gateway pages to databases because these are typically pages made up of input fields and other controls. These pages are *formatted* with HTML and look like any other Web page that uses interactive forms. Behind the scenes, however, are the knobs, dials, and switches that provide access to the actual contents of the database, which are literally incomprehensible to a search engine crawler.

Although these interfaces provide powerful tools for a human searcher, they act as roadblocks for a search engine spider. Essentially, when an indexing spider comes across a database, it's as if it has run smack into the entrance of a massive library with securely bolted doors. A crawler can locate and index the library's address, but because the crawler cannot penetrate the gateway it can't tell you anything about the books, magazines, or other documents it contains.

These Web-accessible databases make up the lion's share of the Invisible Web. They are accessible *via* the Web but may or may not actually be *on* the Web. To search a database you must use the powerful search and

retrieval tools offered by the database itself. The advantage to this direct approach is that you can use search tools that were specifically designed to retrieve the best results from the database. The disadvantage is that you need to find the database in the first place, a task the search engines may or may not be able to help you with.

There are several different kinds of databases used for Web content, and it's important to distinguish between them. Just because Web content is stored in a database doesn't automatically make it part of the Invisible Web. Indeed, some Web sites use databases not so much for their sophisticated query tools, but rather because database architecture is more robust and makes it easier to maintain a site than if it were simply a collection of HTML pages.

One type of database is designed to deliver tailored content to individual users. Examples include My Yahoo!, Personal Excite, Quicken.com's personal portfolios, and so on. These sites use databases that generate "on the fly" HTML pages customized for a specific user. Since this content is tailored for each user there's little need to index it in a general-purpose search engine.

A second type of database is designed to deliver streaming or real-time data—stock quotes, weather information, airline flight arrival information, and so on. This information isn't necessarily customized, but it is stored in a database due to the huge, rapidly changing quantities of information involved. Technically, much of this kind of data is indexable because the information is retrieved from the database and published in a consistent, straight HTML file format. But because it changes so frequently, and has value for such a limited duration (other than to scholars or archivists), there's no point in indexing it. It's also problematic for crawlers to keep up with this kind of information. Even the fastest crawlers revisit most sites monthly or even less frequently (other than news crawlers, which are designed to track rapidly changing news sites). Staying current with real-time information would consume so many resources it is effectively impossible for a crawler.

The third type of Web-accessible database is optimized for the data it contains, with specialized query tools designed to retrieve the information using the fastest or most effective means possible. These are often "relational" databases that allow sophisticated querying to find data that are "related" based on criteria specified by the user. The only way of accessing content in these types of databases is by directly interacting with the database. It is this content that forms the core of the Invisible Web.

Let's take a closer look at these elements of the Invisible Web and demonstrate exactly why search engines can't or won't index them.

Why Search Engines Can't See the Invisible Web

Text—more specifically *hyper*text—is the fundamental medium of the Web. The primary function of search engines is to help users locate

hypertext documents of interest. Search engines are highly tuned and optimized to deal with text pages and, even more specifically, text pages that have been encoded with the HyperText Markup Language (HTML). As the Web evolves and additional media become commonplace, search engines will undoubtedly offer new ways of searching for this information. But for now, the core function of most Web search engines is to help users locate text documents.

HTML documents are simple. Each page has two parts: a "head" and a "body" which are clearly separated in the source code of an HTML page. The head portion contains a title, which is displayed (logically enough) in the title bar at the very top of a browser's window. The head portion may also contain some additional metadata describing the document, which can be used by a search engine to help classify the document. For the most part, other than the title, the head of a document contains information and data that helps the Web browser display the page but is irrelevant to a search engine. The body portion contains the actual document itself. This is the meat that the search engine wants to digest.

The simplicity of this format makes it easy for search engines to retrieve HTML documents, index every word on every page, and store them in huge databases that can be searched on demand. Problems arise when content doesn't conform to this simple Web page model. To understand why, it's helpful to consider the process of crawling and the factors that influence whether a page either can or will be successfully crawled and indexed.

The first thing a crawler attempts to determine is whether access to pages on a server it is attempting to crawl is restricted. Webmasters can use three methods to prevent a search engine from indexing a page. Two methods use blocking techniques specified in the *Robots Exclusion Protocol* that most crawlers voluntarily honor and one creates a technical roadblock that cannot be circumvented (*Robots Exclusion Protocol,* n.d.).

The Robots Exclusion Protocol is a set of rules that enable a Webmaster to specify which parts of a server are open to search engine crawlers, and which parts are off-limits. The Webmaster simply creates a list of files or directories that should not be crawled or indexed and saves this list on the server in a file named robots.txt. This optional file, stored by convention at the top level of a Web site, is nothing more than a polite request to the crawler to keep out, but most major search engines respect the protocol and will not index files specified in robots.txt.

The second means of preventing a page from being indexed works in the same way as the robots.txt file, but it is page-specific. Webmasters can prevent a page from being crawled by including a "noindex" metatag instruction in the "head" portion of the document. Either robots.txt or the noindex metatag can be used to block crawlers. The only difference between the two is that the noindex metatag is page specific, while the

robots.txt file can be used to prevent indexing of individual pages, groups of files, or even entire Web sites.

Password protecting a page is the third means of preventing it from being crawled and indexed by a search engine. This technique is much stronger than the first two since it uses a technical barrier rather than a voluntary standard.

Why would a Webmaster block crawlers from a page using the Robots Exclusion Protocol rather than simply password protecting the pages? Password protected pages can be accessed only by the select few users that know the password. Pages excluded from engines using the Robots Exclusion Protocol, on the other hand, can be accessed by anyone *except* a search engine crawler. The most common reason Webmasters block content from indexing is that a page changes far more frequently than the engines can keep up with.

Pages using any of the three methods described above are part of the Invisible Web. In many cases, they contain no technical roadblocks that prevent crawlers from spidering and indexing the page. They are part of the Invisible Web because the Webmaster has opted to keep them out of the search engines.

Once a crawler has determined whether it is permitted access to a page, the next step is to attempt to fetch it and hand it off to the search engine's indexer component. This crucial step determines to a large degree whether a page is visible or invisible. Let's examine some variations crawlers encounter as they discover pages on the Web, using the same logic they do to determine whether a page is indexable or not.

Case 1

The crawler encounters a page that is straightforward HTML text, possibly including basic Web graphics. This is the most common type of Web page. It is visible and can be indexed, assuming the crawler can discover it.

Case 2

The crawler encounters a page made up of HTML, but it's a form, consisting of text fields, check boxes, or other components requiring user input. It might be a sign-in page, requiring a user name and password. It might be a form requiring the selection of one or more options. The form itself, since it's made up of simple HTML, can be fetched and indexed. But the content *behind* the form (what the user sees after clicking the submit button) may be invisible to a search engine. There are two possibilities here:

- The form is used simply to select user preferences. Other pages on the site consist of straightforward HTML that can be crawled and indexed (presuming there are links from other pages elsewhere on the Web

pointing to the pages). In this case, the form and the content behind it are visible and can be included in a search engine index. Quite often, sites like this are specialized search sites for specific types of content. A good example is Hoover's Business Profiles, which provides a form to search for a company, but presents company profiles in straightforward HTML that can be indexed (*Hoover's Online,* n.d.).

- The form is used to collect user-specified information that will generate dynamic pages when the information is submitted. In this case, although the form is visible, the content "behind" it is invisible. Since the only way to access the content is by using the form, how can a crawler, which is simply designed to request and fetch pages, possibly know what to enter into the form? Since forms can literally have infinite variations, if they function to access dynamic content they are essentially roadblocks for crawlers. A good example of this type of Invisible Web site is the World Bank Group Economics of Tobacco Control Country Data Report Database, which allows you to select any country and choose a wide range of reports for that country (*Economics of Tobacco-Country Data Report,* n.d.). It's interesting to note here that this database is just one part of a much larger site, the bulk of which is fully visible. So even if the search engines do a comprehensive job of indexing the visible part of the site, this valuable information still remains hidden to all but those searchers who visit the site and discover the database on their own.

In the future, forms will pose less of a challenge to search engines. Several projects are underway aimed at creating more intelligent crawlers that can fill out forms and retrieve information. One approach uses preprogrammed "brokers" designed to interact with the forms of specific databases. Other approaches combine brute force with artificial intelligence to "guess" what to enter into forms, allowing the crawler to "punch through" the form and retrieve information. It's not a trivial problem: In a conversation with Google's Chief Technology Officer, Craig Silverstein, he estimated that it may take as long as fifty years before Google has the capability to index all Invisible Web content. And even if general-purpose search engines do acquire the ability to crawl content in databases, it's likely that the native search tools provided by each database will remain the best way to interact with most databases.

Case 3

The crawler encounters a dynamically generated page assembled and displayed on demand. The telltale sign of a dynamically generated page is the "?" symbol appearing in its URL. Technically, these pages are part of the visible Web. Crawlers can fetch any page that can be displayed in a Web browser, regardless of whether it's a static page stored on a server or generated dynamically. A good example of this type of Invisible Web site is

Compaq's experimental SpeechBot search engine, which indexes audio and video content using speech recognition and converts the streaming media files to viewable text (*SpeechBot,* n.d.). Somewhat ironically, one could make a good argument that *most* search engine result pages are *themselves* Invisible Web content, since they generate dynamic pages on the fly in response to user search terms.

Dynamically generated pages pose a challenge for crawlers. Dynamic pages are created by a *script,* a computer program that selects from various options to assemble a customized page. Until the script is actually run, a crawler has no way of knowing what it will actually do. The script *should* simply assemble a customized Web page. Unfortunately, unethical Webmasters have created scripts to generate literally millions of similar but not quite identical pages in an effort to "spamdex" the search engine with bogus pages. Sloppy programming can also result in a script that puts a spider into an endless loop, repeatedly retrieving the same page.

These "spider traps" can be a real drag on the engines, so most have simply made the decision not to crawl or index URLs that generate dynamic content. They're "apartheid" pages on the Web—separate but equal, making up a big portion of the "opaque" Web that potentially can be indexed but is not. Inktomi's FAQ about its crawler, named "Slurp," offers this explanation:

> Slurp now has the ability to crawl dynamic links or dynamically generated documents. It will not, however, crawl them by default. There are a number of good reasons for this. A couple of reasons are that dynamically generated documents can make up infinite URL spaces, and that dynamically generated links and documents can be different for every retrieval so there is no use in indexing them. (*Slurp,* n.d.)

As crawler technology improves, it's likely that one type of dynamically generated content will increasingly be crawled and indexed. This is content that essentially consists of static pages that are stored in databases for production efficiency reasons. As search engines learn which sites providing dynamically generated content can be trusted not to subject crawlers to spider traps, content from these sites will begin to appear in search engine indices. It's important to note that even as search engines learn which content is acceptable, they still may not index everything, as evidenced by this statement from Google's Webmaster tips page: "We are able to index dynamically generated pages. However, because our web crawler can easily overwhelm and crash sites serving dynamic content, we limit the amount of dynamic pages we index" (*Google Information for Webmasters,* n.d.).

Another development that has reduced the barriers for dynamic content is the increasing adoption of *paid inclusion* programs by the major search engines. These programs are designed to allow Webmasters to specify specific pages for crawling and guaranteed indexing, in exchange for an annual fee. The search engines give no preferential treatment to these

pages beyond guaranteed indexing, and spam rules still apply. Any pages that violate search engine spam policies, whether crawled or submitted via paid exclusion, are subject to removal from the index. Paid inclusion is a means for search engines to trust dynamic content, on the theory that nobody would willingly pay just to have their content removed anyway.

Case 4

The crawler encounters an HTML page with nothing to index. There are thousands, if not millions, of pages that have a basic HTML framework, but which contain only Flash; images in the .gif, .jpeg, or other Web graphics format; streaming media; or other nontext content in the body of the page. These types of pages are truly parts of the Invisible Web because there's nothing for the search engine to index. Specialized multimedia search engines are able to recognize some of these nontext file types and index minimal information about them, such as file name and size, but these are far from keyword searchable solutions.

Case 5

The crawler encounters a site offering dynamic, real-time data. There are a wide variety of sites providing this kind of information, ranging from real-time stock quotes to airline flight arrival information. These sites are also part of the Invisible Web, because these data streams are, from a practical standpoint, unindexable. While it's technically possible to index many kinds of real-time data streams, the value would only be for historical purposes, and the enormous amount of data captured would quickly strain a search engine's storage capacity, so it's a futile exercise. A good example of this type of Invisible Web site is Cheap Ticket's FlightTracker, which provides real-time flight arrival information taken directly from the cockpit of in-flight airplanes (*FlightTracker,* n.d.).

Case 6

The crawler encounters a PDF or Postscript file. PDF and Postscript are text formats that preserve the look of a document and display it identically regardless of the type of computer used to view it. While many search engines index PDF files, most do not index the full text of the documents. Google stops indexing after 120KB; AlltheWeb stops indexing after 110KB.

An experimental search engine called ResearchIndex, created by computer scientists at the NEC Research Institute, not only indexes the full text of PDF and Postscript files, it also takes advantage of the unique features that commonly appear in documents using the format to improve search results (*CiteSeer,* n.d.). For example, academic papers typically cite other documents and include lists of references to related material. In addition to indexing the full text of documents, ResearchIndex also creates a citation index that makes it easy to locate related documents. It also appears

that citation searching has little overlap with keyword searching, so combining the two can greatly enhance the relevance of results.

Case 7

The crawler encounters a database offering a Web interface. There are tens of thousands of databases containing extremely valuable information available via the Web. But search engines cannot index the material in them. Although we present this as a unique case, Web-accessible databases are essentially a combination of cases 2 and 3. Databases generate Web pages dynamically, responding to commands issued through an HTML form. Though the interface to the database is an HTML form, the database itself may have been created before the development of HTML, and its legacy system is incompatible with protocols used by the engines, or they may require registration to access the data. Finally, they may be proprietary, accessible only to select users, or users who have paid a fee for access.

Ironically, the original HTTP specification developed by Web inventor Tim Berners-Lee included a feature called format negotiation that allowed a client to say what kinds of data it could handle and allow a server to return data in any acceptable format. Berners-Lee's vision encompassed the information in the Invisible Web, but this vision, at least from a search engine standpoint, has largely been unrealized.

These technical limitations give you an idea of the problems encountered by search engines when they attempt to crawl Web pages and compile indices. There are other, nontechnical reasons why information isn't included in search engines. We look at those next.

Four Types of Invisible

Technical reasons aside, there are other reasons that some kinds of material that can be accessed either on or via the Internet are not included in search engines. There are really four "types" of invisible Web content. We make these distinctions not so much to make hard and fast distinctions between the types, but rather to help illustrate the amorphous boundary of the Invisible Web that makes defining it in concrete terms so difficult.

The four types of invisible are:

- The "Opaque" Web,
- The Private Web,
- The Proprietary Web, and
- The Truly Invisible Web.

The "Opaque" Web

The "Opaque" Web consists of files that *can be,* but are not, included in search engine indices. The Opaque Web is quite large and presents a unique challenge to a searcher. Whereas the deep content in many truly

Invisible Web sites is accessible if you know how to find it, material on the Opaque Web is often much harder to find.

The biggest part of the Opaque Web consists of files that the search engines *can* crawl and index, but simply do not. There are a variety of reasons for this; let's look at them.

Depth of Crawl

Crawling a Web site is a resource-intensive operation. It costs money for a search engine to crawl and index every page on a site. In the past, most engines would merely sample a few pages from a site rather than performing a "deep crawl" that indexed every page, reasoning that a sample provided a "good enough" representation of a site that would satisfy the needs of most searchers. Limiting the depth of crawl also reduced the cost of indexing a particular Web site.

In general, search engines don't reveal how they set the depth of crawl for Web sites. Increasingly, there is a trend to crawl more deeply, to index as many pages as possible. As the cost of crawling and indexing goes down, and the size of search engine indices continues to be a competitive issue, the depth of crawl issue is becoming less of a concern for searchers. Nonetheless, simply because one, fifty, or five thousand pages from a site are crawled and made searchable, there is no guarantee that every page from a site will be crawled and indexed. This problem gets little attention and is one of the top reasons why useful material may be all but invisible to those who only use general-purpose search tools to find Web materials.

Frequency of Crawl

The Web is in a constant state of dynamic flux. New pages are added constantly, and existing pages are moved or taken off the Web. Even the most powerful crawlers typically visit only about 10 million pages per day, a fraction of the entire number of pages on the Web. This means that each search engine must decide how best to deploy its crawlers, creating a schedule that determines how frequently a particular page or site is visited.

Web Search researchers Steve Lawrence and Lee Giles, writing in the July 8, 1999, issue of *Nature,* state that "indexing of new or modified pages by just one of the major search engines can take months" (Lawrence and Giles, 1999). While the situation appears to have improved since their study, most engines only completely "refresh" their indices monthly or even less frequently.

It's not enough for a search engine to simply visit a page once and then assume it's still available thereafter. Crawlers must periodically return to a page to not only verify its existence, but also to download the freshest copy of the page and perhaps fetch new pages that have been added to a site. According to one study, it appears that the half-life of a Web page is some-

what less than two years and the half-life of a Web site is somewhat more than two years. Put differently, this means that if a crawler returned to a site spidered two years ago it would contain the same number of URLs, but only half of the original pages would still exist, having been replaced by new ones ("Graph Structure in the Web," n.d.; "Altavista, Compaq, and IBM," n.d.).

New sites are the most susceptible to oversight by search engines because relatively few other sites on the Web will have linked to them compared to more established sites. Until search engines index these new sites, they remain part of the Invisible Web.

Maximum Number of Viewable Results

It's quite common for a search engine to report a very large number of results, sometimes into the millions of documents. However, most engines also restrict the total number of results they will display for a query, typically between 200 and 1,000 documents. For queries that return a huge number of results, this means that the majority of pages the search engine has determined might be relevant are inaccessible, since the result list is arbitrarily truncated. Those pages that don't make the cut are effectively invisible.

Good searchers are aware of this problem and will take steps to circumvent it by using a more precise search strategy and the advanced filtering and limiting controls offered by many engines. However, for many inexperienced searchers this limit on the total number of viewable hits can be a problem. What happens if the answer you need is available (with a more carefully crafted search) but cannot be viewed using your current search terms?

Disconnected URLs

For a search engine crawler to access a page, one of two things must take place. Either the Web page author uses the search engine's "Submit URL" feature to request that the crawler visit and index the page, or the crawler discovers the page on its own by finding a link to the page on some other page. Web pages that aren't submitted directly to the search engines, and that don't have links pointing to them from other Web pages, are called "disconnected" URLs and cannot be spidered or indexed simply because the crawler has no way to find them.

Quite often, these pages present no technical barrier for a search engine. But the authors of disconnected pages are clearly unaware of the requirements for having their pages indexed. A May 2000 study by IBM, AltaVista, and Compaq discovered that the total number of disconnected URLs makes up about 20 percent of the potentially indexable Web, so this isn't an insignificant problem ("Graph Structure in the Web," n.d.; "Altavista, Compaq, and IBM," n.d.).

In summary, the Opaque Web is large, but is not impenetrable. Determined searchers can often find material on the Opaque Web, and search engines are constantly improving their methods for locating and indexing Opaque Web material.

The three other types of invisible are more problematic, as we'll see.

The Private Web

The Private Web consists of technically indexable Web pages that have deliberately been excluded from inclusion in search engines. There are three ways Webmasters can exclude a page from a search engine:

- Password protect the page. A search engine spider cannot go past the form that requires a username and password;
- Use the robots.txt file to disallow a search spider from accessing the page;
- Use the "noindex" metatag to prevent the spider from reading past the head portion of the page and indexing the body.

For the most part, the Private Web is of little concern to most searchers. Private Web pages simply use the public Web as an efficient delivery and access medium, but in general are not intended for use beyond the people who have permission to access the pages.

There are other types of pages that have restricted access that may be of interest to searchers, yet they typically aren't included in search engine indices. These pages are part of the "Proprietary" Web, which we describe next.

The Proprietary Web

Search engines cannot for the most part access pages on the Proprietary Web, because these pages are only accessible to people who have agreed to special terms in exchange for viewing the content. Proprietary pages may simply be content that's only accessible to users willing to register to view them. Registration in many cases is free, but a search crawler clearly cannot satisfy the requirements of even the simplest registration process.

Other types of proprietary content are available only for a fee, whether on a per-page basis or via some sort of subscription mechanism. Examples of proprietary fee-based Web sites include Hoover's and the Wall Street Journal Interactive Edition.

Proprietary Web services are not the same as traditional online information providers, such as Dialog, Lexis-Nexis, and Dow Jones. These services offer Web access to proprietary information but use legacy database systems that existed long before the Web came into being. While the con-

tent offered by these services is exceptional, they are not considered to be Web or Internet providers.

The Truly Invisible Web

Some Web sites or pages are truly invisible, meaning that there are technical reasons that search engines can't spider or index the material they have to offer. A definition of what constitutes a truly invisible resource must necessarily be somewhat fluid, since the engines are constantly improving and adapting their methods to embrace new types of content. But at the time of writing truly invisible content consisted of several types of resources.

The simplest, and least likely to remain invisible over time, are Web pages that use file formats that current generation Web crawlers aren't programmed to handle. These file formats include PDF, postscript, Flash, Shockwave, executables (programs), and compressed files. There are two reasons search engines do not currently index these types of files. First, the files have little or no textual context, so it's difficult to categorize them, or compare them for relevance to other text documents. The addition of metadata to the HTML container carrying the file could solve this problem—but it would nonetheless be the metadata description that got indexed rather than the contents of the file itself.

The second reason certain types of files don't appear in search indices is simply because the search engines have chosen to omit them. They *can* be indexed, but aren't. You can see a great example of this in action with the Research Index engine, which retrieves and indexes PDF, Postscript, and even compressed files in real time, creating a searchable database that's specific to your query. AltaVista's Search Engine product for creating local site search services is capable of indexing more than 250 file formats, but the flagship public search engine includes only a few of these formats. It's typically lack of willingness, not an ability issue with file formats.

More problematic are dynamically generated Web pages. Again, in some cases, it's not a technical problem but rather unwillingness on the part of the engines to index this type of content. This occurs specifically when a noninteractive script is used to generate a page. These are static pages, and generate static HTML that the engine could spider. The problem is that unscrupulous use of scripts can also lead crawlers into "spider traps" where the spider is literally trapped within a huge site of thousands, if not millions, of pages designed solely to spam the search engine. This is a major problem for the engines, so they've simply opted not to index URLs that contain script commands.

Finally, information stored in relational databases, which cannot be extracted without a specific query to the database, is truly invisible.

Crawlers aren't programmed to understand either the database structure, or the command language used to extract information.

Conclusion

The Invisible Web is a vast portion of cyberspace, and offers invaluable resources that should not be overlooked by serious searchers. Although search engine technology continues to improve, the Invisible Web is largely an intractable problem that will be with us for some time to come. Although it's a vast and useful resource, it's important not to get bogged down in the semantics. An information professional should treat these types of resources like traditional reference tools. Learn what's available and have them ready to go. The best way for searchers to access the Invisible Web is to build and bookmark a personal collection of resources, treating them as a personal "reference library," and using them when needed, rather than relying on search engines that in many cases simply cannot access the content residing on the Invisible Web.

References

Altavista, Compaq, and IBM researchers create world's largest, most accurate picture of the Web. (n.d.). [Summary of "Graph Structure in the Web," (n.d.)]. Retrieved August 27, 2003, from http://www.almaden.ibm.com/almaden/webmap_release.html.

CiteSeer: The NEC Research Institute Scientific Literature Digital Library. (n.d.). Retrieved April 17, 2003, from http://www.researchindex.com. Commonly referred to as ResearchIndex.

Economics of tobacco-country data report. (n.d.). Retrieved April 14, 2003, from http://www1.worldbank.org/tobacco/database.asp.

FlightTracker. (n.d.). Retrieved April 16, 2003, from CheapTickets Travel Web site: http://www.cheaptickets.com/trs/cheaptickets/flighttracker/flight_tracker_graphic.xsl.

Google catalogs. (n.d.). Retrieved April 17, 2003, from http://catalogs.google.com.

Google information for Webmasters. (n.d.). Retrieved April 17, 2003, from http://www.google.com/webmasters/2.html.

Graph structure in the Web. (n.d.). Retrieved August 27, 2003, from http://www9.org/w9cdrom/160/160.html.

Hoover's online. (n.d.). Retrieved April 15, 2003, from http://www.hoovers.com/.

Lawrence, S., & Giles, C. L. (1999). Accessibility of information on the web. *Nature, 400,* 107–109. Additional material can be found at: http://wwwmetrics.com/.

Robots exclusion. (n.d.). Retrieved April 17, 2003, from http://www.robotstxt.org/wc/exclusion.html.

Shadow TV. (n.d.). Retrieved March 17, 2003, from http://www.shadowtv.com.

Singingfish. (n.d.). Retrieved April 9, 2003, from http://www.singingfish.com/.

Slurp. (n.d.). Retrieved April 17, 2003, from http://www.inktomi.com/slurp.html.

Speechbot. (n.d.). Retrieved April 15, 2003, from http://www.speechbot.com.

Additional Readings

Guernsey, L. (2001, January 25). Mining the "deep web" with sharper shovels. *New York Times,* p. G1.

Price, G. (2002). Specialized search engine FAQs: More questions, answers, and issues. *Searcher, 10*(9), 42–48.

Price, G., & Sherman, C. (2001). Exploring the invisible Web: Seven essential strategies. *Online, 25*(4), 32–35.

Sherman, C. (2001). Google unveils more of the invisible Web. *Search Day.* Retrieved April 17, 2003, from http://www.searchenginewatch.com/searchday/article.php/2158091.

Web Search: Emerging Patterns

AMANDA SPINK

ABSTRACT

THIS ARTICLE EXAMINES the public searching of the Web and provides an overview of recent research exploring what we know about how people search the Web. The article reports selected findings from studies conducted from 1997 to 2002 using large-scale Web user data provided by commercial Web companies, including Excite, Ask Jeeves, and AlltheWeb.com. We examined what topics people search for on the Web; how people search the Web using keywords in queries during search sessions; and the different types of searches conducted for multimedia, medical, e-commerce, sex, etc., information. Key findings include changes and differences in search topics over time, including a shift from entertainment to e-commerce searching by largely North American users. Findings show little change in current patterns of Web searching by many users from short queries and sessions. Alternatively, we see more complex searching behaviors by some users, including successive and multitasking searches.

INTRODUCTION

People are spending increasing amounts of time working with electronic information. Web searching services such as Alta Vista and Google are now everyday tools for information seeking.

The research that explores such issues as the organization of the Web or Web searching trends is becoming more important for users and Web search engines alike. There are many overlapping and related levels of a user's context that are relevant to Web research, including the information environment/social level, organizational level, information-seeking

Amanda Spink, Associate Professor, School of Information Sciences, University of Pittsburgh, 610 IS Building, 135 N. Bellefield Avenue, Pittsburgh PA 15260.
LIBRARY TRENDS, Vol. 52, No. 2, Fall 2003, pp. 299–306

level, human-computer interaction level, and query level. In order to better understand how to organize the Web, we also need to understand more about how people interact with and use the Web at these different levels.

For many users, Web interactions are often frustrating and constrained. A growing body of large-scale quantitative or qualitative studies is exploring these issues, including the effectiveness and limitations of Web search engines (Lawrence & Giles, 1998) and how users search the Web (Silverstein et al., 1999; Wolframn et al., 2001). One outgrowth of Web research is better support for human information behaviors and the development of a new generation of Web tools, such as Web meta–search engines, to help users persist in electronic information seeking and help people resolve their information problems.

This article reports selected results from a large-scale and ongoing series of studies of searching behavior on commercial Web search engines by a diverse range of users. The research reported in this article is focused at the human-computer interaction and query level of Web user behavior. Selected results are reported from studies of Web query data from Excite, AlltheWeb.com, and Ask Jeeves. The researchers were not able to obtain data from the major Web company Google, but further analysis is being conducted on Web query data from Alta Vista. The goal of these studies is to track trends in the public searching of the Web and explore how the public searches the Web (Spink, Wolfram, Jansen, & Saracevic, 2001).

Web Query Data Sets

The analysis was conducted on various large sets of Web query data provided by various Web companies from 1997 to 2001. All users were anonymous and could not be identified in any way. But we could identify each user's sequence of queries.

Each transaction record contained three fields. With these three fields, researchers were able to locate a user's initial query and recreate the chronological series of actions by each user in a session:

Time of Day: measured in hours, minutes, and seconds;
User Identification: an anonymous user code assigned by the Web server;
Query Terms: exactly as entered by the given user.

We focused on three levels of data analysis—*sessions, queries,* and *terms.* This large-scale study provides insights into Web searching with implications for developing better search engines and services.

Web Search Patterns

Selected findings, summarized below, provide interesting insights into current patterns of public Web searching, including how people structure

their Web searches, what they search for, and search behavior in special topic areas.

Web Queries

How long are general Web queries? The mean length of Excite queries increased steadily from 1.5 words in 1996 to 2.6 in 1999, and the mean number of terms in unique queries was 2.4. The mean query length for U.S./U.K. users in 1996 was 1.5 and mean query length for European users in 1997 was 1.5—in 1999 U.S./U.K. users mean query length was 2.6, and for European users it was 1.9. English language queries increased in length more quickly than European language queries. Jansen, Spink, and Saracevic (2000) report that Web queries were short and most users did not enter many queries per search. The mean number of queries per user was 2.8 in 1997.

However, a sizable percentage of users did go on to either modify their original query or view subsequent results. On average, a query contained 2.21 terms in 1997. About one in three queries had one term only, two in three had one or two terms, and four in five had one, two, or three terms. Fewer than 4 percent of the queries were comprised of more than six terms. Spink, Jansen, Wolfram, and Saracevic (2002) reported the mean terms per query had increased slightly to 2.6 by 2001. Overall, general Web queries are still short.

Use of Boolean Operators

How frequently are Boolean operators used during Web searching? The use of Boolean operators (AND, OR, NOT, +, -) increased from 22 percent of queries in 1997 to 28 percent of queries in 1999. From the 1996–99 data set, approximately 8 percent of searches included proximity searching. Jansen, Spink, and Saracevic (2000) found that Boolean operators were seldom used. One in eighteen users used any Boolean capabilities and, of the users employing them, every second user made a mistake, as defined by Excite rules. The '+' and '-' modifiers that specify the mandatory presence or absence of a term were used more than Boolean operators. About one in twelve users employed them. About one in eleven queries incorporated a '+' or '-' modifier. But a majority of these uses were mistakes (about two out of three). Spink, Jansen, Wolfram, and Saracevic (2002) reported that by 2001 some 10 percent of Web searches contained Boolean operators. Overall, we see that Boolean search is still in limited use.

Web Query Reformulation

Do Web search engine users reformulate their queries? Spink, Jansen, and Ozmultu (2000) found that most users searched one query only and did not follow with successive queries. The average session, ignoring identical queries, included 1.6 queries. About two in three users submitted a

single query, and six in seven did not go beyond two queries. Spink, Jansen, Wolfram, and Saracevic (2002) reported that in 2001 some 44 percent of users modified their queries with 25 percent of users entering three or more queries. Overall, most users still enter only one or two queries and conduct little query reformulation.

Question and Request Format Web Queries

Do users enter queries in question or request format? Spink and Ozmutlu (2002) report that only 50 percent of Ask Jeeves users entered queries in question format. Most questions began with the words "Where do I find . . . ?" Some 25 percent of users phrased their queries as requests, most commonly "Get me information. . . ." Overall, most general Web queries are in query rather than question format.

Search Terms: Distribution

What is the distribution of search terms? Jansen, Spink, and Saracevic (2000) report the distribution of the frequency of use of terms in queries as highly skewed. A few terms were used repeatedly and many terms were used only once. On the top of the list, the sixty-three subject terms that had a frequency of appearance of 100 or more represented only one-third of 1 percent of all terms, but they accounted for about one of every ten terms used in all queries. Terms that appeared only once amounted to half of the unique terms. By 2001, 615 terms were not repeated in the dataset, as reported by Spink, Jansen, Wolfram, and Saracevic (2002). Overall, Web searching involves a small percentage of high-frequency terms and many low-frequency terms.

Use of Relevance Feedback

How frequently are relevance feedback commands used? Analysis of Web searches shows that, when available, relevance feedback is rarely used. About one in twenty queries used the feature "More Like This." Spink, Jansen, and Ozmultu (2000) found that one-third of Excite users went beyond the single query, with a smaller group using either query modification or relevance feedback or viewing more than the first page of results. They examined the occurrence of each query type (unique, modified, relevance feedback, view a results page, etc.) in a large sample of user sessions. The distribution of query type changes as the length of the user session increases. For the user sessions of two and three queries, the relevance feedback query is dominant. As the length of the sessions increase, the occurrences of relevance feedback as a percentage of all query types decreases. Some 63 percent of relevance feedback sessions could be construed as being successful. If the partially successful user sessions are included, then more than 80 percent of the relevance feedback sessions provided some measure of success.

Viewing Results

How many pages of ten hits do users view? This is a very interesting question for users and Web industry people alike. From 1996 to 1999, for more than 70 percent of the time, a user only viewed the top ten results. On average, users viewed 2.35 pages of results (where one page equals ten hits). Over half the users did not access results beyond the first page. Jansen, Spink, and Saracevic (2000) found that more than three in four users did not go beyond viewing two pages. By 2001, only roughly one-third of users looked beyond the second page of Web sites retrieved (Spink, Jansen, Wolfram, & Saracevic, 2002).

Web Search Topics

Users search the Web on an infinite variety of topics. The next section focuses on what we know about how users search on particular topics such as sex, e-commerce, and medical information. Spink, Jansen, Wolfram, and Saracevic (2002) report a shift in Web search topics from entertainment and sex in 1997 to commerce, travel, employment, economy, people, places, and things in 2001. Search topics have shifted from entertainment to e-commerce as the content of the Web has shifted more toward business.

Sexually Related Searching

Jansen, Spink, and Saracevic (2000) found searching about sex on Excite represents only a small proportion of all searches. When the top frequency terms are classified as to subject, the top category is "Sexual." As to the frequency of appearance, about one in every four terms in the list of sixty-three highest used terms can be classified as sexual in nature. But while sexual terms are high as a category, they still represent a very small proportion of all terms. Many other subjects are searched and the diversity of subjects searched is very high.

Spink, Ozmultu, and Lorence (in press) found that sexually related searches were longer than general searches and involved viewing more pages of Web sites. Overall, sexual Web searchers are more persistent and likely to be seeking images.

Medical and Health-related Web Searching

Medical and health-related information is proliferating on the Web. Spink, Yang, Nykanen, Lorence, Ozmutlu, and Ozmutlu (in press) found that a small percentage of Web searching is medical or health-related. The top five categories of medical or health advice sought were general health, weight issues, reproductive health and puberty, pregnancy/obstetrics, and human relationships. Trends show that medical and health queries have declined as a proportion of Web queries as the use of specialized medical/health Web sites and e-commerce–related queries has increased, but e-commerce–related searching has increased substantially.

E-Commerce Searching

E-commerce queries are increasing on the Web (Spink & Guner, 2001). Web queries are a primary means for translating people's business product, service, and information needs for e-commerce. Spink and Guner (2001) found that business queries often include more search terms than other types of queries, are less modified, lead to fewer Web pages viewed, and include less advanced search features. Company or product name queries were the most common form of business. The most common business-related query submitted to Ask Jeeves was "Where can I buy . . . " or the request "I want to buy . . . " Spink, Jansen, Wolfram, and Saracevic (2002) found that by 2001 the largest category of Web searches were e-commerce related.

Multimedia Searching

Goodrum and Spink (2001) conducted a specific analysis of image queries within the 1.2 million queries. Provisions for image searching by Web search engines are important for users. Users seeking images input relatively few terms to specify their image information needs on the Web. Users seeking images interact iteratively during the course of a single session but input relatively few queries overall. Most image terms are used infrequently with the top term occurring in less than 9 percent of queries.

Jansen, Spink, and Saracevic (2000) found that many terms were unique in the large data sets, with over half of the terms used only once. Terms indicating sexual or adult content materials appear frequently in image queries. They represented a quarter of the most frequently occurring terms but were a small percentage of the total terms. Overall, multimedia searching is shifting as the content of the Web changes (Jansen, Goodrum, & Spink, 2000; Ozmutlu, Spink, & Ozmutlu, 2002).

Longitudinal Search Patterns

Despite the generally short nature of user Web queries and search sessions, recent studies are also showing that some users are engaging in more complex Web search interactions.

Successive Searching

How many Web searches do users conduct on a particular topic? Spink, Bateman, and Jansen (1999) conducted an interactive survey of over three hundred Excite users and found that many had conducted two searches or three or more related searches using the Excite search engine over time when seeking information on a particular topic. Successive searches often involved a refinement or extension of the previous searches as new databases were searched and search terms changed as the Excite users' understanding and evaluation of results evolved over time from one successive search to the next.

Multitasking Search

How many topics are users searching for? Spink, Ozmutlu, and Ozmutlu (2002) found that many Web searches involved users seeking information on two or more topics concurrently. Overall, we see some users moving toward more complex searches that involve multiple related interactions and multiple topics.

Discussion

The research we conducted over the last five years shows some interesting patterns and trends in general Web searching. In summary, most Web queries are short, without much modification, and simple in structure. Few queries incorporate advanced search techniques and, when they are used, many mistakes result. However, advanced search features are slowly growing in use. Many people retrieve a large number of Web sites, but view few results pages and tend not to browse beyond the first or second results pages. Overall, a small number of terms are used with high frequency and many terms are used once. Web queries are very rich in subject diversity, and some are unique. The subject distribution of Web queries does not seem to map to the distribution of Web sites' subject content. Some users are engaging in more longitudinal Web searching practices during their information-seeking processes that are not well supported by Web search technologies. We can see that Web searching is growing as a huge public challenge, but it is an imprecise and challenging skill.

Insights into Web searching trends and patterns have implications for the organization of the Web. A key problem for Web organization is that people in general do not really understand how Web search engines work or the structure of the Web. The Web is a creature of interaction, yet many Web interactions are subject to limitations due to a lack of information and training by users. In general, Web search engines do not explain the Web to users and do not tell users that their search engines only cover a limited number of Web sites. Web culture is based on a "quick and dirty" approach to searching, rather than an exploratory, interactive approach. Web organizational issues and search issues are related. The success of users' search interactions depends on the intersection of more effective search techniques and self–user training.

Conclusion and Further Research

Our ongoing study of Web searching is examining a number of large-scale Web query transaction logs. These studies, using large-scale log data, are showing some interesting trends and patterns in general Web searching and helping to answer some interesting questions about Web searching. Due to the nature of the data, the research cannot address the results of users' queries or assess the performance of different search engines. However, the findings do provide a snapshot for comparison of public Web

searching that can help improve Web search engines and services. Further research is currently being conducted, using query data from Alta Vista, to explore Web search including the similarities and/or differences between North American and European users. Ongoing Web user behavior research is further identifying trends and impacting the development of new types of user training, interfaces and software agents, and new organizational schemas to aid users in better Web searching.

REFERENCES

Goodrum, A., & Spink, A. (2001). Image searching on the Excite web search engine. *Information Processing and Management, 37*(2), 95–312.

Jansen, B. J., Goodrum, A., & Spink, A. (2000). Searching for multimedia: An analysis of audio, video, and image Web queries. *World Wide Web: An International Journal, 3*(4), 249–254.

Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real life, real users and real needs: A study and analysis of users' queries on the Web. *Information Processing and Management, 36*(2), 207–227.

Lawrence, S., & Giles, C. L. (1998). Searching the World Wide Web. *Science, 280*(5360), 98–100.

Ozmutlu, S., Spink, A., & Ozmutlu, H. C. (2002). Trends in multimedia Web searching: 1997–2001. *Information Processing and Management, 38*(3), 475–496.

Silverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1999). Analysis of a very large Web search engine query log. *ACM SIGIR Forum, 33*, 3.

Spink, A., Bateman, J., & Jansen, B. J. (1999). Searching the Web: Survey of EXCITE users. *Internet Research: Electronic Networking Applications and Policy, 9*(2), 117–128.

Spink, A., & Guner, O. (2001, July). E-commerce Web queries: Excite and Ask Jeeves study. *First Monday, 6*(7).

Spink, A., Jansen, B. J., & Ozmultu, H. C. (2000). Use of query reformulation and relevance feedback by Web users. *Internet Research: Electronic Networking Applications and Policy, 10*(4), 317–328.

Spink, A., Jansen, B. J., Wolfram, D., & Saracevic, T. (2002). From e-sex to e-commerce: Web search changes. *IEEE Computer, 35*(3), 133–135.

Spink, A., & Ozmutlu, H. C. (2002). Characteristics of question format Web queries: An exploratory study. *Information Processing and Management, 38*(4), 453–471.

Spink, A., Ozmutlu, H. C., & Lorence, D. P. (in press). Web searching for sexual information: An exploratory study. *Information Processing and Management.*

Spink, A., Ozmutlu, H. C., & Ozmutlu, S. (2002). Multitasking information seeking and searching processes. *Journal of the American Society for Information Science and Technology, 53*(8), 639–652.

Spink, A., Wolfram, D., Jansen, B. J., & Saracevic, T. (2001). Searching the Web: The public and their queries. *Journal of the American Society for Information Science, 53*(2), 226–234.

Spink, A., Yang, Y., Nykanen, P., Lorence, D. P., Jansen, B. J., Ozmutlu, S., & Ozmutlu, H. C. (in press). Medical and health Web searching: An exploratory study.

Wolfram, D., Spink, A., Jansen, B. J., & Saracevic, T. (2001). Vox populi: The public searching of the Web. *Journal of the American Society for Information Science and Technology, 52*(12), 1073–1074.

Copyright Law and Organizing the Internet

REBECCA P. BUTLER

ABSTRACT

UNITED STATES INTELLECTUAL PROPERTY LAW, specifically that covering copyright, has important implications for American libraries. This article considers the following: fair use and the Internet; current and prospective law and electronic media, especially concerning interlibrary loan and online reserves; publishers and users; and the impact that copyright law has on the role of the library and the issue of free access.

INTRODUCTION

Did you know that every e-mail you write, every Web page you create, anything that you generate in a tangible form is automatically copyrighted by United States law—whether you officially register it with the U.S. Copyright Office or not (Bruwelheide, 1995, p. 7)? Because the readers of *Library Trends* tend to be those of us associated with libraries, probably, yes, you do know this. Yet copyright law, especially that associated with electronic communications, continues to be a quagmire from which it is difficult to extract oneself, one's employment environment (library), or one's patrons.

Copyright is a serious matter that carries implications for organizing the Internet from both the viewpoints of the owners and publishers of a work to the work's users. This article will discuss several strands within the dilemma of the Internet and copyright: the law, including fair use; public domain; the Digital Millennium Copyright Act; the Technology, Education, and Copyright Harmonization Act; the Sonny Bono Extension Act; owners and users of copyrighted works and how the library and its role with

Rebecca P. Butler, Associate Professor, Educational Technology, Research and Assessment, Northern Illinois University, DeKalb, IL 60115
LIBRARY TRENDS, Vol. 52, No. 2, Fall 2003, pp. 307–317

respect to access comes into play; what we as librarians can do to make intellectual property a "smoother sell" to those with whom we work; and other intellectual property issues that may impact our interpretations of copyright law. Indeed, copyright law and organizing the Internet is a conundrum.

Current Law

Below is a discussion of some important areas (for those of us in libraries) of the current copyright law, along with examples.[1]

Please note that, in reference to this article, all examples will include some use of the Internet.

Fair Use

Those who work with and/or study copyright are well aware of the vagueness within the law. It is never more clear/unclear than when determining how much one can reproduce from a copyrighted work before being considered in violation of copyright law. Section 107 of the 1976 Copyright Act states that the amount of material we borrow from a copyrighted work depends on four factors:

- Purpose and character of use,
- Nature of the work,
- Part being copied, and
- Work's marketability.[2]

These four fair use factors must *all* be in place for a portion of an item to be considered to fall under fair use restrictions.

The first of the four fair use factors, purpose and character of use, covers what the borrower wants to do with the copied material. "Copying for nonprofit, educational, or personal reasons leans in favor of fair use . . ." (Butler, 2001, p. 35). Thus, if you are an academic librarian, sending a personal e-mail with a paragraph from *Statistical Methods for the Social and Behavioral Sciences* (Marascuilo & Serlin, 1988) to a group of interested statistics students, you should be all set with factor one!

The second fair use factor, nature of the work, deals with the characteristics of the work one wishes to copy; in other words, "whether the work is fact or fiction, published or unpublished" (Butler, 2001, p. 35). Nonfiction and published media is most likely to fit this second factor. Thus, the academic librarian above is still in compliance, since *Statistical Methods* is nonfiction and was published in the 1980s.

The part of the work being copied, the third fair use factor, is a little more subject to debate. While the less amount one copies, the better, this factor is measured both quantitatively and qualitatively. Generally speaking, quantity is based on how much needs to be copied to achieve the objec-

tive and how much such an amount is in comparison to the total of the original. In addition, there is the issue of quality. Here the "heart" of the work comes into play. The heart of a work can vary from a tiny slice of an item to a huge portion. Therefore, if one sentence is the "heart" of *Statistical Methods,* copying it can be in violation (Butler, 2001, p. 35). Luckily for our academic librarian above, this does not seem to be the case with this particular copying example.

The fourth fair use factor is concerned with the marketability of the work, should copying of it occur. Chances are that e-mailing a group of college statistics students a paragraph out of *Statistical Methods* will not affect the sales of this book negatively, so here again our academic librarian is probably safe.

Remember, if you are not sure if you are in copyright law compliance when borrowing, it is still best to contact the owner of the work for permission.[3]

Sonny Bono Copyright Term Extension Act (CTEA)

In an effort to maintain consistency between the United States and the other members of the Berne Convention,[4] in 1998 Congress passed the Sonny Bono Copyright Term Extension Act (CTEA). So named because Congressman Bono was working on this at the time of his death, CTEA extends the duration of copyright in the United States retroactively from the life of the author plus fifty years to the life of the author plus seventy years and, in the case of works for hire and those under a corporate ownership, from seventy-five to ninety-five years or one hundred twenty years (whichever comes first) (Hoffman, 2001; Wikipedia, 2003b). This act was challenged as unconstitutional (*Eldred v. Ashcroft*) and brought before the Supreme Court in 2002. In January 2003, the Supreme Court found it constitutional (Wikipedia, 2003a). For libraries providing information via the Internet, CTEA means that we will have to get copyright clearance for much of what our patrons request for a longer period of time.

Public Domain

If all materials created were in the public domain, there would be no need for copyright law and litigation. Public domain determines that the owner of a copyrighted work has given up that ownership to the general public to use in any way that it pleases. Thus, someone creating a Web page can access a public domain electronic clip art Web site; borrow a graphic; place this on another Web page; modify the object in size, or color, or by adding or subtracting characteristics, whatever—all without worrying about obtaining permission to borrow or create a derivative work.[5]

Media in the public domain does not need to state that it is so. However, without that statement, interpreting whether or not an item is in the public domain is a somewhat complicated activity. General interpretation

is that content is in the public domain if it was published before 1923. Additionally, any work created after January 1, 1978, will be in copyright until seventy years after the death of the last author or one hundred twenty years from the date of creation (in the case of works-for-hire) (Gasaway, 2001; Project Gutenberg, 2002). For the years in between, things can get a little complicated. For example, works published between 1964 and 1977, if they have a copyright notice, have a twenty-eight–year copyright term with an automatic extension of sixty-seven more years. An informative table, entitled "When Works Pass into the Public Domain," by Lolly Gasaway is available at http://www.unc. edu/~unclung/public-d.htm. It explains the various rules of public domain in regard to the year an item was published.

For those librarians concerned with electronic interlibrary loan and online reserves, public domain can be a wonderful thing. There is then no need to search for owners of works, ask for permissions, etc.

Digital Millennium Copyright Act (DMCA)

According to Gretchen McCord Hoffman in *Copyright in Cyberspace: Questions and Answers for Librarians* (2001), the Digital Millennium Copyright Act (DMCA) has far-reaching effects on copyright law, in a number of areas ranging from electronic communications to international copyright law, to exemptions for library reproductions, to anticircumvention technologies, to distance education, among others. For example, the DMCA can provide protection for libraries that are online service providers (OSP) in the instance of copyright violations, if the library/provider registers an agent and develops policies for notification and termination of the service use should copyright violations be discovered (Hoffman, 2001).

For those of us concerned with libraries and the Internet, the DMCA is difficult to summarize, and several articles could be written in this area alone. It is possible that this law may end up influencing libraries in such arenas as "services, research, website development, distance education, and Internet access" (Crews, 2000, p. 116). For the purposes of this particular article, the points below illustrate some of the ways that the DMCA may affect those of us in libraries in terms of the Internet.

Given the WIPO Copyright Treaties section of the DMCA,

- "[C]opyright owners [can] impose technological controls and other restrictions on the use of their works, and . . . constrain the use of materials for research and teaching in a manner more restrictive than may be established under existing copyright law" (Crews, 2000, p. 117). Thus, the owner of a copyrighted Web page could attach charges or restrictions to the use of his/her work by a library, even if the library's use was under fair use.

- "[T]he restrictions [in the first point above] may not apply to particular classes of works and to particular persons, if the restrictions would 'adversely affect' the ability to make 'noninfringing uses' of those works, as determined by the U.S. Copyright Office";
- Libraries may circumvent protections if they are evaluating a work for prospective purchase;
- Every three years, the Librarian of Congress will "conduct proceedings to examine and review the effect of the restrictions on the availability and use of copyrighted works, especially for education and libraries";
- Reverse engineering and encryption research of software may take place in libraries (Crews, 2000, p. 117).

Under the Online Service Provider Liability section of the DMCA, libraries (if they are an OSP):

- May not be held liable for copyright infringement committed by those using their online services;
- Must remove or disable access to infringing media;
- Must adopt a policy terminating the service of those users who do not abide by copyright law;
- Need to designate an agent to deal with copyright infringements (Crews, 2000, p. 118).

Because there is so much in the DMCA that can influence libraries in terms of the Internet, whether it is as an online service provider, use of interlibrary loan, distance education, etc., it is best to study the DMCA to determine where it effects your specific library setting and how.[6]

Technology, Education, and Copyright Harmonization (TEACH) Act

While the TEACH Act only indirectly affects most libraries, it is mentioned here due to its currency as one of the newest of our copyright laws. In effect, the TEACH Act gives institutional users (faculty, staff, and students) more rights to use and borrow materials for use in distance education than those previously provided under the 1976 copyright law. The TEACH Act, which became law in the later part of 2002, provides for fair use portions of a variety of instructional works in a distance education setting, if the providing institution follows a number of rules. These rules include that the institution is educational, nonprofit and accredited; works copied are lawfully obtained; materials are required for instruction, etc. Prior to its passage, remote classrooms, such as those connected through online education, television, and other means, had very few rights in comparison to face-to-face classrooms. The TEACH Act does not provide for "digital delivery of supplemental reading materials"

(Harper, 2002). Thus, in libraries, we will still need to abide by the fair use guidelines.

UCITA

In early 2003, the American Bar Association rejected the Uniform Computer Information Transaction Act (UCITA, 2003). While this is good news for libraries nationwide, the fact that UCITA is actually contract law means that every state has the option of whether to pass it or not. Currently, only two states, Maryland and Virginia, have passed UCITA. UCITA is a controversial and confusing law for those of us in libraries and for our patrons. It is a threat to the fair use doctrine as it applies to electronic media in that it validates both shrink and click-wrap licenses and replaces copyright law with contract law, thus allowing users to "click away their fair use rights" (Hoffman, 2001, p. 55; Kunze, 2000). In simple terms, the passage of UCITA in a particular state could mean that a library which owns a book and computer software purchased as a package might find itself being able to lend out the book but *not* the corresponding software. This could affect not only regular library circulation but also interlibrary loan: "its consequence would be to preempt copyright law with un-negotiated contract law, that is, to replace user rights under the copyright law, such as fair use, with agreements to give up those rights that users never have the opportunity to negotiate" (Hoffman, 2001, p. 147).

Other

Currently in the House and Senate a variety of federal copyright legislation is waiting for discussion, support, passage into law, etc. While these acts represent possible future legislation, we only look in this article to copyright law as it currently stands. However, be aware of new legislation as it occurs by keeping current in professional print journals; online sources, such as those provided by the ALA Web site; and professional organizations.

Owners and Users: Libraries and Internet Access

Interlibrary Loan (ILL)

Because no library is able to have everything that its patrons will need at every point in time, interlibrary loan (ILL) is necessary. In the past this involved either copying print material or sending the original and mailing these items to the receiving library through snail mail. In the world of electronic communications, ILL can be much quicker (just scan and send the item electronically) and much more difficult (how do the sending and receiving libraries work within copyright guidelines, since electronic copying and sending may mean a minimum of four extra copies being available at one time)?[7]

[The use of a fax or scanner to transmit copyrighted material is resolved if the library only uses these tools as transmission devices] (Martin, 2003).

Section 108 of the U.S. Copyright Law works with ILL within certain lender and borrower limits. For example, a lending library may send only one copy of one article from a specific journal or periodical issue. (Patrons who wish for a copy of more than one article from a specific journal should purchase the journal issue, subscribe to the journal, or pay a copyright fee.) Additionally, the Commission on New Technological Uses of Copyrighted Works (CONTU) Interlibrary Loan Guidelines state that a patron may borrow up to five copies of articles (but no more) from a specific journal within a given year. Therefore, whether the transmission is electronic or print, there is a limit to how much may be borrowed. In addition, the borrowing library must certify that its requests fit within the CONTU guidelines and must retain records to that effect for at least three years. It is also possible for ILL to occur through the use of URLs to access posted articles or databases for which a library has a license (Besenjak, 1997, pp. 156–157; Martin, 2003).

Online Reserves

Electronic reserves are often viewed with unease by those working in academic, special, public, and many other library settings due to the potential for unrestricted access, piracy, and violation of the fourth fair use factor (marketability). While print reserves in many libraries are traditionally offered under the fair use guidelines, electronic reserves may provide users with the ability to transmit copies to others as well as printout copies of works that are still under copyright law. What does this mean for libraries—those providers of reserves? Well, it actually means a great deal. The library involved needs to develop a series of checks and balances to protect itself from copyright violation and litigation and to protect its users as well. This is done in a number of ways. What libraries should do in terms of materials, electronic reserve, and copyright is described below. Libraries need to:

- Check that the material they put on reserve, which is not owned by them (for example, it may have been provided by an instructor), has been obtained in a lawful manner;
- Obtain appropriate permissions, if necessary;
- Pay royalties as needed;
- Follow the fair use guidelines, if no permission has been sought;
- Limit access;[8]
- Put on reserve as little an amount of the material as is feasible to satisfy course and user needs;
- Include a reference section and copyright notice from the original work on the electronic reserve item;
- Keep works on electronic reserve as short a time as possible (for example, one semester per class);
- Avoid putting problem items on electronic reserve;[9]

- Limit use of audio and video streaming;[10]
- Link to databases, instead of scanning items, if library licenses or subscriptions permit this;
- Remove access to the work once the course is over (Hoffman, 2001; Martin, 2003).

Thus, online reserves can be an easy way to provide materials to patrons—under the proper guidelines.

More Intellectual Property Issues to Consider

Two other issues involving intellectual property and the Internet require some time at this point: privacy and piracy. While not actually part of copyright law, these two areas are influential to libraries working in an electronic, i.e., Internet environment. They are briefly covered below.

Privacy

"These . . . illustrate the dilemma librarians face in protecting patron confidentiality. Because one reads controversial literature does not necessarily mean that one is a threat to national security or society" (Weiner, 1997). "Broadly defined, privacy is regarded as information about oneself that is kept from others. In the library setting, right to privacy refers to the lack of availability of information about oneself" (Winter, 1997). What this means, in the library/information-seeking setting, is that a patron's personal circulation records, online reserves, reference questions, Internet access, and interlibrary loan requests should not be available to more than those library personnel who need the information in order to provide the patron with what s/he needs. As early as 1939, the American Library Association "recognized the right to privacy . . . in its Code of Ethics for Librarians" (Mitchell, 2003b). Mitchell (2003b) also states that "the right to privacy in a library is also implicit in the ALA's Library Bill of Rights, which guarantees free access to library resources for all users and opposes any limitations on the right to an individual's exercise of free expression. . . . Through the Library Bill of Rights and the Code of Ethics, librarians fight to protect patron privacy and preserve our democratic society."

The potential for concerns with privacy issues comes from three major areas: "1) protecting libraries records; 2) making patrons aware of records that others can create based on their interactions while on library computers or networks; and 3) requiring vendor partners to adhere to an appropriate level of privacy protection" (Mitchell, 2003a). Perhaps because new legislations, such as the U.S.A. Patriot Act, may disagree with the idea of privacy in American libraries, abuse to privacy appears to be growing. Thus, privacy remains another topic, besides copyright law, which affects patrons' use of and access to library materials.

Piracy

Piracy is a term used to identify the unlawful use or borrowing of a copyrighted item. It is in violation of copyright law. The term is most often used in connection with the illegal use or copying of computer software and is considered a felony (Simpson, 2001, pp. 9, 84). It can also be applied to such things as intercepting satellite video transmissions (p. 74). Like privacy, piracy is not actually regulated by copyright law, but it is a closely related issue.

What Should We Do?

Given the discussion above, what should we do when confronted with Internet copyright infringement by our colleagues and/or clientele? Certainly, neither pointing out their lack of integrity or ignorance in obeying copyright law will be popular stances. However, there are some ways to make such instances win/win or, at least, learn/learn situations. We, as the copyright experts, can:

- Educate our audience through such venues as copyright workshops, in-services, classes, DVDs, videos, and teleconferences;
- Keep abreast of the most current changes in the law;[11]
- Be available for consultation by patrons and colleagues;
- Obtain support from those in our organization's administration;
- Be calm and understanding when confronting an infringement;
- Encourage correct action;
- Give examples of libelous actions and responses by the law to such actions;
- Cite law;
- Encourage users to read documentation;
- Encourage citing of information obtained from another source;
- Retain an intellectual properties attorney;
- Use original sources;
- Demonstrate ethical behavior;
- Remind our colleagues and clientele that we are all liable for our own actions.

Conclusion

In terms of copyright violations, ignorance is not bliss. This message alone is worth repeating to those who assume that because they are 1. educators, 2. not copying "much," 3. unable to find the owner of the work, 4. and other excuses, that they are not "really" in violation and/or will "never" get caught. While there are no "copyright police" commonly running from library to library, there are people willing to report violations *and* companies willing to pay for these reports (Butler, 2002, p. 42). Thus, it is imperative that those of us in libraries, whether working with patrons or behind the scenes, abide by copyright law.

NOTES

1. Caveat: This article covers the U.S. law and issues in using Internet materials published in the United States. International issues are a separate subject and will not be addressed here, unless needed for definition purposes.
2. Although for this article's purposes we are addressing the Internet, this discussion applies to all sorts of media, not just those accessible via the Internet.
3. Copyright permission information is not included in this article. A wide variety of print and electronic sources are available on this subject, however. The following references support such information: *About SESAC,* retrieved August 15, 2001, from http://www.sesac.com/aboutsesac/aboutsesac1.html; American Society of Composers, Authors, and Publishers, retrieved June 26, 2001, from http://www.ascap.com/licensing/about.html; *ASCAP Licensing: Frequently Asked Questions About Licensing,* (2001), retrieved June 25, 2001, from http://ascap.com/licensing/licensingfaz.html; Association of American Publishers, *How to Request Copyright Permission from Publishers,* (1998), retrieved June 25, 2001, from http://www.publishers.org/home/about a/highered/howtopg.htm; BMI and Performing Rights, retrieved August 15, 2001, from http://www.bmi.com/licensing/; Brad Templeton, (n.d.), *10 Big Myths About Copyright Explained,* retrieved June 25, 2001, from http://www.templetons.com/brad/copymyths.html; J. H. Bruwelheide, (1995), *The Copyright Primer for Librarians and Educators,* 2nd ed. (Chicago: American Library Association); *Clearing the Way for Your Rights,* (1999), retrieved June 26, 2001, from http://www.presentations.com/create/organiz/1999/06/31_fl_cop_04.html; *Copyright Permission Letter,* (1996), retrieved June 26, 2001, from http://www.bham.wednet.edu/copyperm.htm; *Fair Use: Obtaining Permissions,* Georgia Harper, (1997), *Sample Letter Requesting Permission,* retrieved June 1, 2001, from http://www.utsystem.edu/ogc/intellectualproperty/permmm.htm; Fulcrun Publishing, (n.d.), *How to Apply for Permission,* retrieved June 25, 2001, from http://fulcrum-books.com/html/permissions.html; "Getting Permission," retrieved June 25, 2001, from http://www.utsystem.edu/ogc/intellectualproperty/permissn.htm; Illinois Association of School Boards, (1999, February), *General Personnel: Exhibit—Request to Reprint Material,* 5.170-E: 1; Motion Picture Licensing Corporation, retrieved August 15, 2001, from http://www.mplc.com/index2.htm; "Organizing Your Message: Getting Copyright Permission," retrieved June 26, 2001, from http://www.presentations.com/create/organiz/1999/06/31_fl_cop_04.html; O'Reilly & Associates, Inc., (2001), *Permission Guidelines,* retrieved June 25, 2001, from http://www. oreilly.com/oreilly/author/permission/; R. S. Talub, (2001, May/June), "Permissions, 'Fair Use', and Production Resources for Educators and Librarians, Part I of II," *TechTrends, 45*(3), 8; *Requesting Permission,* (n.d.), retrieved June 25, 2001, from http://depts.Washington.edu/uwcopy/use/obtainingrights/5.shtml; "Zip Through Permissions as Never Before—Over the Web!" retrieved June 26, 2001, from http://www.copyright.com.
4. The Berne Convention is one of two major international copyright treaties (the other is the Universal Copyright Convention) to which the United States adheres. Because there is no common copyright law in the world, these two conventions' members agree to abide by and give each other the same copyright protection that is given in their own countries (Besenjak, 1997, p. 48).
5. This is, of course, assuming that the clip art site really is in public domain. It is possible for a site creator to claim that all clip art (or other works) are in public domain when, in fact, some or all of these items are borrowed from copyrighted sites (Butler, 2000).
6. The Digital Millennium Copyright Act can be found in full at http://www.loc.gov/copyright/legislation/dmca.pdf. Another helpful site for libraries in terms of the DMCA and the Sonny Bono Copyright Term Extension Act is Arnold P. Lutzker's site, entitled "Primer on the Digital Millennium: What the Digital Millennium Copyright Act and the Copyright Term Extension Act Mean for the Library Community." It is found at http://www.arl.org/info/frn/copy/primer.html.
7. These extra copies could be 1. the copy scanned from the print version and placed on the hard drive of the library providing the copy; 2. the copy on the hard drive of the receiving library's computer; 3. the copy sent onward electronically to the patron; and 4. the copy the patron prints off of his/her computer.

8. Electronic access can be limited in a number of ways. For example, a library system may require a password from the user to enter the online reserve area. Access can also be limited with the use of class membership lists and/or retrieval by course number or the instructor's name (Martin, 2003).
9. "Problem" items might include student papers, unpublished pieces, course packs, textbooks, sample tests, etc. (Martin, 2003).
10. According to Charlie Morris in "Streaming Audio," http://wdvl.com/Multimedia/Sound/Audio/streaming.html, audio and video streaming occur when audio and video files are able to play on your computer while you are still downloading them.
11. Professional journals and Internet sites for major library organizations, such as the American Library Association, represent excellent ways to remain current.

REFERENCES

Besenjak, C. (1997). *Copyright plain & simple.* Franklin Lakes, NJ: Career Press.

Bruwelheide, J. H. (1995). *The copyright primer for librarians and educators.* 2nd ed. Chicago: American Library Association.

Butler, R. P. (2000). Copyright and computers 2000. St. Charles, IL: Illinois Computing Educators Annual Meeting.

Butler, R. P. (2001). Copyright as a social responsibility: Fair use—I need it now! *Knowledge Quest, 29,* 35–36.

Butler, R. P. (2002). Software piracy: Don't let it byte you! *Knowledge Quest, 31,* 41–42.

Copyright law of the United States of America and related laws contained in Title 17 of the United States code. (2001). Circular 92.

Crews, K. D. (2000). *Copyright essentials for librarians and educators.* Chicago: American Library Association.

Gasaway, L. (2001). When works pass into the public domain. Retrieved February 10, 2003, from http://www.unc.edu/~unclung/public-d.htm.

Harper, G. (2002). The TEACH Act finally becomes law. Retrieved March 17, 2003, from bttp://www.utsystem.edu/ogc/intellectualproperty/teachact.htm.

Hoffman, G. M. (2001). *Copyright in cyberspace: Questions and answers for librarians.* New York: Neal-Schuman Publishers.

Kunze, C. A. (2000). UCITA online: Uniform computer information transactions act. Retrieved March 19, 2003, from http://www.ucitaonline.com/.

Library of Congress. (2003). United States copyright office. Retrieved March 10, 2003, from http://www.loc.gov/copyright/.

Marascuilo, L. A., & Serlin, R. C. (1998). *Statistical methods for the social and behavioral sciences.* New York: W. H. Freeman.

Martin, R. (2003). The library and copyright issues. Presentation to ETT 590: Instructional Technology Workshop: Copyright. DeKalb, IL: Northern Illinois University.

Mitchell, K. (2003a). Privacy tutorial. Message #2. Chicago: American Library Association.

Mitchell, K. (2003b). Privacy tutorial. Message #4. Chicago: American Library Association.

Project Gutenberg Official Home Site. (2002). Public domain and copyright how-to. Retrieved February 10, 2003, from http://www.gutenberg.net/vol/pd.html.

Simpson, C. (2001). *Copyright for schools: A practical guide,* 3rd ed. Worthington, OH: Linworth Publishing.

UCITA Presentation. (2003). The "new" UCITA for 2003—Chapter relations/COL briefing. Philadelphia: American Library Association.

Weiner, R. G. (1997). Privacy and librarians: An overview. Retrieved March 11, 2003, from http://www.txla.org/pubs/tlj-1q97/privacy.html.

Wikipedia. (2003a). Eldred v. Ashcroft. Retrieved March 17, 2003, from http://www.wikipedia.org/wiki/Eldred_v_Ashcroft.

Wikipedia. (2003b). Sonny Bono Copyright Term Extension Act. Retrieved March 17, 2003, from http://www.wikipedia.org/wiki/Sonny_Bono_Copyright_Term_Extension_Act.

Winter, K. A. (1997). Privacy and the rights and responsibilities of librarians. Retrieved March 11, 2003, from http://alexia.lis.uiuc.edu/review/winter1997/winter.html.

A Survey of Metadata Research for Organizing the Web

JANE L. HUNTER

ABSTRACT

THIS ARTICLE ATTEMPTS TO PROVIDE an overview of the key metadata research issues and the current projects and initiatives that are investigating methods and developing technologies aimed at improving our ability to discover, access, retrieve, and assimilate information on the Internet through the use of metadata.

1. INTRODUCTION

The rapid expansion of the Internet has led to a demand for systems and tools that can satisfy the more sophisticated requirements for storing, managing, searching, accessing, retrieving, sharing, and tracking complex resources of many different formats and media types.

Metadata is the value-added information that documents the administrative, descriptive, preservation, technical, and usage history and characteristics associated with resources. It provides the underlying foundation upon which digital asset management systems rely to provide fast, precise access to relevant resources across networks and between organizations. The metadata required to describe the highly heterogeneous, mixed-media objects on the Internet is infinitely more complex than simple metadata for resource discovery of textual documents through a library database. The problems and costs associated with generating and exploiting such metadata are correspondingly magnified.

Metadata standards, such as Dublin Core, provide a limited level of interoperability between systems and organizations to enable simple resource discovery. But, there are still many problems and issues that remain

Jane L. Hunter, Distributed Systems Technology Centre Pty. Ltd., Level 7, GP South, University of Queensland, St. Lucia, Queensland 4072, Australia
LIBRARY TRENDS, Vol. 52, No. 2, Fall 2003, pp. 318–344

to be solved. Cory Doctorow (2001) believes that the vision of an Internet in which everyone describes their goods, services, or information using concise, accurate, and common or standardized metadata that is universally understood by both machines and humans is a "pipe-dream, founded on self-delusion, nerd hubris and hysterically inflated market opportunities." Other people cite the popularity and efficiency of Google as an example of an extremely successful search engine that does not depend on expensive and unreliable metadata. Google combines PageRanking (in which the relative importance of a document is measured by the number of links to it) with sophisticated text-matching techniques to retrieve precise, relevant, and comprehensive search results (Brin & Page, 1998).

Some of the major disadvantages of metadata are cost, unreliability, subjectivity, lack of authentication, and lack of interoperability with respect to syntax, semantics, vocabularies, languages, and underlying models. However, there are many researchers currently investigating strategies to overcome different aspects of these limitations in an effort to provide more efficient means of organizing content on the Internet. Other researchers are investigating metadata to describe the new types of real-time streaming content being generated by emerging broadband and wireless applications to enable both push and pull of this content based on users' needs. The goal of this article is to provide an overview of some of the key metadata research underway that is expected to improve our ability to search, discover, retrieve, and assimilate relevant information on the Internet regardless of the domain or format.

2. The Key Research Areas

In this section I have identified what I consider to be some of the key metadata research areas, both now and over the next few years. The following subsections provide a brief description of the work being undertaken and some key citations for each of the research areas summarized in the list below:

- Extensible Markup Language (XML)—XML and its associated technologies—XML Namespaces, XML Query languages, and XML Databases—are enabling implementers to develop metadata application profiles (XML Schemas) that combine metadata terms from different namespaces to satisfy the needs of a particular community or application. Large-scale XML descriptions of content are being stored in XML Databases and can be queried using XML Query Language. These are key technologies to enabling the automated computer processing, integration, and exchange of information.
- Semantic Web technologies—"The Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation" (Berners-Lee,

Hendler, & Lassila, 2001). There are two main building blocks for the semantic Web:

  - Formal languages—RDF (Resource Description Framework), DAML+OIL, and OWL (Web Ontology Language), which is being developed by the Web Ontology Working Group of the W3C.
  - Ontologies—communities will use the formal languages to define both domain-specific ontologies and top-level ontologies to enable relationships between ontologies to be determined for cross-domain searching, exchange, and information integration.
- Web Services—using open standards such as WSML, UDDI, and SOAP, Web services will enable the building of software applications without having to know who the users are, where they are, or anything else about them.
- Metadata Harvesting—the Open Archives Initiative (OAI) provides a protocol for data providers to make their metadata and content accessible—enabling value-added search and retrieval services to be built on top of harvested metadata.
- Multimedia metadata—there will be a further move away from textual resources to new multimedia formats that support better quality and higher compression ratios, e.g., images (JPEG-2000), video (MPEG-4), audio (MP3), 3D (VRML, Web3D), multimedia (SMIL, Shockwave Flash), and interactive digital objects. All of these new media types will require complex fine-grained metadata, extracted automatically where possible.
- Rights metadata—new emerging standards such as MPEG-21 and XrML are designed to enable automated copyright management and services.
- Automatic metadata extraction—technologies to enable the automatic classification and segmentation of digital resources. In particular, automatic image processing, speech recognition, and video-segmentation tools will enable content-based querying and retrieval of audiovisual content.
- Search engines:
  - Smarter agent-based search engines;
  - Federated search engines;
  - Peer-to-peer search engines;
  - Multimedia search engines;
  - Multilingual search engines;
  - New search interfaces—search interfaces that present results graphically;
  - Automatic/dynamic aggregation and generation of search results into hypermedia and multimedia presentations.
- Personalization/customization—autonomous agents that push relevant information to the user based on user preferences that may be personally configured or learned by the system.

- Broadband networks—multigigabit-capable networks for high-quality video-conferencing and visualization applications:
  - Grid computing—distributed computing and communications infrastructures for data intensive computing applications;
  - The Semantic Grid—the combination of semantic Web technologies with grid computing to provide large scale data access and integration to the e-Science community.
- Mobile and wireless technologies—delivery of information to mobile devices or appliances based on users' current context or location.
- Authentication—technologies to ensure trust and record the provenance of metadata.
- Annotation systems—enable users to attach their own subjective notes, opinions, and views to resources for others to access and read.
- Preservation metadata—metadata to support long-term preservation strategies for all types of digital resources.

2.1 XML Technologies and Metadata

XML and its associated technologies—XML Namespaces, XML Query languages, and XML Databases—are enabling implementers to develop metadata schemas, application profiles, large repositories of XML metadata, and search interfaces using XML Query Language. These technologies are key to enabling the automated computer-processing, integration, and exchange of information over the Internet.

2.1.1 Extensible Markup Language (XML). XML (W3C XML, 2003) is a simple, very flexible text format derived from SGML (ISO 8879). Originally designed to meet the challenges of large-scale electronic publishing, XML is playing an increasingly important role in the exchange of a wide variety of data on the Web and elsewhere. Because XML makes it possible to exchange data in a standard format, independent of storage, it has become the de facto standard for representing metadata descriptions of resources on the Internet.

2.1.2 XML Schema Language. XML Schema Language (W3C XML Schema, 2003) provides a means for defining the structure, content, and semantics of XML documents. It provides an inventory of XML markup constructs, which can constrain and document the meaning, usage, and relationships of the constituents of a class of XML documents: datatypes, elements and their content, attributes and their values, entities and their contents, and notations. Thus, the XML Schema Language can be used to define, describe, and catalog XML vocabularies for classes of XML documents, such as metadata descriptions of Web resources or digital objects.

XML Schemas have been used to define metadata schemas for a number of specific domains or applications—such as METS (Library of Congress, 2003), MPEG-7 (Martinez, 2002), MPEG-21 (Bormans & Hill, 2002), and NewsML (IPTC, 2001). An additional major metadata development

has been the employment of W3C's XML Schemas and XML Namespaces to combine metadata elements from different domains/namespaces into "application profiles" or metadata schemas that have been optimized for a particular application. For example, a particular community may want to combine elements of Dublin Core (DCMI, 2003), MPEG-7 (Martinez, 2002), and IMS (IMS, 2003) to enable the resource discovery of audio-visual learning objects.

2.1.3 XML Query. The mission of the XML Query Working Group (W3C XML Query, 2003) is to provide flexible query facilities to extract data from real and virtual documents on the Web, thereby providing the needed interaction between the Web world and the database world. Ultimately, collections of XML files will be accessed like databases. The new query language, XQuery, is still evolving, but it will provide a functional language comprised of several kinds of expressions that can be nested or composed with full generality. A working draft version of XQuery and a list of current XQuery implementations is available at http://www.w3.org/XML/Query.html.

2.1.4 XML Databases. There is a large amount of research and development going on in the area of XML databases. Ronald Bourret provides an excellent overview of the current state of this work and a comparison of current XML database technologies (Bourret, 2003a; Bourret, 2003b). Bourret divides XML Database solutions into the following categories:

- Middleware—software you call from your application to transfer data between XML documents and databases;
- XML-enabled databases—databases with extensions for transferring data between XML documents and themselves;
- Native XML databases—databases that store XML in "native" form, generally as some variant of the DOM mapped to an underlying data store. This includes the category formerly known as persistent DOM (PDOM) implementations;
- XML servers—XML-aware J2EE servers, Web application servers, integration engines, and custom servers. Some of these are used to build distributed applications while others are used simply to publish XML documents to the Web. Includes the category formerly known as XML application servers;
- Content Management Systems (CMS)—applications built on top of native XML databases and/or the file system for content/document management and which include features such as check-in/check-out, versioning, and editors;
- XML query engines—standalone engines that can query XML documents;
- XML data binding—products that can bind XML documents to objects. Some of these can also store/retrieve objects from the database.

2.1.5 Metadata Schema Registries. A number of groups have been tackling the issue of establishing registries of metadata schemas to enable the reuse and sharing of metadata vocabularies and to facilitate semantic interoperability. In particular the CORES project (CORES, 2003), which builds on the work of SCHEMAS (SCHEMAS, 2002), is exploring the use of metadata schema registries in order to enable the reuse of existing schemas, vocabularies, and application profiles that have been "registered."

2.2 The Semantic Web and Interoperability

According to Tim Berners-Lee, director of the World Wide Web Consortium (W3C), the Semantic Web is "an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation.... The Semantic Web will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users" (Berners-Lee, Hendler, & Lassila, 2001). But the Semantic Web has a long way to go before this dream is realized. The real power of the Semantic Web will be realized when programs and applications are created that collect Web content from diverse sources, process the information, and exchange the results with other programs.

Two of the key technological building blocks for the Semantic Web are:

- Formal languages for expressing semantics, such as the Resource Description Framework (RDF), DAML+OIL, and OWL (Web Ontology Language), which have been/are being developed within the W3C's Semantic Web Activity (W3C Semantic Web Activity, 2002); and
- The ontologies that are being constructed from such languages.

2.2.1 Formal Languages: RDF, DAML+OIL, OWL. The general consensus appears to be that while XML documents and schemas are ideal for defining the structural, formatting, and encoding constraints for a particular domain's metadata scheme, a different type of language is required for defining meaning or semantics.

The Resource Description Framework (RDF) (W3C, RDF Syntax, & Model Recommendation, 1999; W3C RDF Vocabulary Description Language, 2003) uses triples to makes assertions that particular things (people, Web pages, or whatever) have properties (such as "is a sister of," "is the author of") with certain values (another person, another Web page). The triples of RDF form webs of information about related things. Because RDF uses URIs to encode this information in a document, the URIs ensure that concepts are not just words in a document but are tied to a unique definition that everyone can find on the Web. This work is being undertaken by the RDF Core Working Group of the W3C.

The W3C Web Ontology Working Group (W3C Web Ontology, 2003) is building upon the RDF Core work to develop a language for defining structured Web-based ontologies that will provide richer integration and interoperability of data among descriptive communities. This is the Web Ontology Language (OWL) (W3C, OWL, 2003), which in turn is building upon the DAML+OIL (DAML+OIL, 2001) specification developed by DARPA.

2.2.2 Ontologies. An ontology consists of a set of concepts, axioms, and relationships that describes a domain of interest. An ontology is similar to a dictionary or glossary but with greater detail and structure and expressed in a formal language (e.g., OWL) that enables computers to process its content. Ontologies can enhance the functioning of the Web to improve the accuracy of Web searches and to relate the information in a resource to the associated knowledge structures and inference rules defined in the ontology.

Upper ontologies provide a structure and a set of general concepts upon which domain-specific ontologies (e.g., medical, financial, engineering, sports, etc.) could be constructed. An upper ontology is limited to concepts that are abstract and generic enough to address a broad range of domain areas at a high level. Computers utilize upper ontologies for applications such as data interoperability, information search and retrieval, automated inferencing, and natural language processing.

A number of research and standards groups are working on the development of common conceptual models (or upper ontologies) to facilitate interoperability between metadata vocabularies and the integration of information from different domains. The Harmony project developed the ABC Ontology/Model (Lagoze & Hunter, 2001)—a top-level ontology to facilitate interoperability between metadata schemas within the digital library domain. The CIDOC CRM (CIDOC CRM, 2003) has been developed to facilitate information exchange in the cultural heritage and museum community. The Standard Upper Ontology (SUO, 2002) is being developed by the IEEE SUO Working Group.

Many communities are developing domain-specific or application-specific ontologies. Some examples include biomedical ontologies such as OpenGALEN (OpenGALEN, 2002) and SNOMED CT (SNOMED CT, 2003), financial, and sporting ontologies such as the soccer, baseball, or running ontologies in the DAML Ontology Library (DAML Ontology Library, 2003).

A large number of research efforts are focusing on the development of tools for building and editing ontologies (Denny, 2002)—these are moving towards collaborative tools such as OntoEdit (Sure et al., 2002) and built-in support for RuleML to enable the specification of inferencing rules.

2.2.4 Topic Maps. Topic Maps (Topic Maps, 2000) is a new ISO standard for a system describing knowledge structures and associating them with

information resources. They provide powerful ways of navigating large and interconnected corpora. Instead of replicating the features of a book index, the topic map generalizes them, extending them in many directions at once. The difference between Topic Maps and RDF is that Topic Maps are centered on topics while RDF is centered on resources. RDF annotates the resources directly whilst topic maps create a "virtual map" above the resources, leaving them unchanged.

2.2.5 Ontology Storage and Querying. A number of research groups are currently working on the development of inferencing tools and deductive query engines to enable the deduction of new information or knowledge from assertions or metadata and ontologies expressed in formal ontology languages (RDF, DAML+OIL, or OWL). A technical report on "Ontology Storage and Querying," published recently by ICS FORTH in Crete, provides a very good survey of the current state of ontology storage and querying tools (Magkanaraki et al., 2002).

2.3 Web Services

Web services (W3C Web Services Activity, 2003) are a relatively new concept, expected to evolve rapidly over the next few years. They could be the first major practical manifestation of Semantic Web–based thinking. Detailed definitions vary, but Web services will enable the building of software applications without having to know who the users are, where they are, or anything else about them. In the next few years, Web services may be developed that can be understood and used automatically by the computing devices of users and of public libraries. External Application Services Providers (ASPs) may also provide such services. Web services are based on open, Internet standards. The core standards and protocols for Web services are being developed and are expected to be finalized by 2003. They include (in addition to XML):

- Web Services Description Language (WSDL) (WSDL, 2003), which enables a common description of Web Services;
- Universal Description, Discovery, and Integration (UDDI) (OASIS, 2003) registries, which expose information about a business or other entity and its technical interfaces;
- Simple Object Access Protocol (SOAP)/XML Protocol (W3C XML Protocol Working Group, 2003), which enables structured message exchange between computer programs.

The concept of Web services is currently being developed under the banner of e-commerce. However, there do appear to be potential applications for public sector service providers. For example, search interfaces could be accessed or provided as Web services by public libraries or by Application Service Providers (ASPs) on their behalf.

2.4 Metadata Harvesting—The Open Archives Initiative (OAI)

The Open Archives Initiative (OAI) (OAI, 2003) is a community that has defined an interoperability framework, the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), to facilitate the sharing of metadata. Using this protocol, data providers are able to make metadata about their collections available for harvesting through an HTTP-based protocol. Service providers then use this metadata to create value-added services. OAI-PMH Version 2.0 was released in February 2003 (OAI-PMH, 2003).

To facilitate interoperability, data providers are required to supply metadata that complies to a common schema, the unqualified Dublin Core Metadata Element Set. Additional schemas are also allowed and are distinguished through the use of a metadata prefix.

Although originating in the E-Print community, OAI data providers now include a number of multimedia collections such as the Library of Congress American Memory collection (Library of Congress, 2002), OpenVideo (OpenVideo, 2002), and University of Illinois historical images (UIL, 2002). DSpace at MIT (DSpace, 2002) is also a registered data provider. HP Labs and MIT Libraries have also made the DSpace software available—it is an open-source, digital asset management software platform that enables institutions to capture and describe digital works using a submission workflow module; distribute an institution's digital works over the Web through a search and retrieval system; and store and preserve digital works over the long term. And it supports OAI-PMH Version 2.0.

To date, OAI service providers have mostly developed simple search and retrieval services (OAI Registered Service Providers, 2002). These include Arc, citebaseSearch, and my.OAI. Scirius searches and retrieves specifically scientific data—from the Web, proprietary databases, and Open Archives. One of the more interesting services is DP9, a gateway service that allows traditional Web search engines (e.g., Google) to index otherwise hidden information from OAI archives. The DSTC's MAENAD project developed a search, retrieval, and presentation system for OAI that searches for and retrieves mixed-media resources on a particular topic, determines the semantic relationships between the retrieved objects, and combines them into a coherent multimedia presentation, based on their relationships to each other (Little, Guerts, & Hunter, 2002).

2.5 Multimedia Metadata

Audiovisual resources in the form of still pictures, graphics, 3D models, audio, speech, and video will play an increasingly pervasive role in our lives and, because of the complex information-rich nature of such content, value-added services such as analysis, interpretation, and metadata creation become much more difficult, subjective, time consuming, and expensive. Audiovisual content requires some level of computational interpretation

and processing in order to generate metadata of useful granularity efficiently. Standardized multimedia metadata representations that will allow some degree of machine interpretation will be necessary. The MPEG-7 and MPEG-21 standards have been developed to support such requirements.

2.5.1 MPEG-7 Multimedia Content Description Interface. MPEG-7 (Martinez, 2002), the "Multimedia Content Description Interface," is an ISO/IEC standard for describing multimedia content, developed by the Moving Pictures Expert Group (MPEG). The goal of this standard is to provide a rich set of standardized tools to enable both humans and machines to generate and understand audiovisual descriptions that can be used to enable fast, efficient retrieval from digital archives (pull applications) as well as filtering of streamed audiovisual broadcasts on the Internet (push applications). MPEG-7 can describe audiovisual information regardless of storage, coding, display, transmission, medium, or technology. It addresses a wide variety of media types including still pictures, graphics, 3D models, audio, speech, video, and combinations of these (e.g., multimedia presentations). The MPEG-7 specification provides:

- A core set of Descriptors (Ds) that can be used to describe the various features of multimedia content;
- Predefined structures of Descriptors and their relationships, called Description Schemes (DSs).

MPEG-7 Multimedia Description Schemes enable descriptions of multimedia content, including:

- Information describing the creation and production processes of the content (director, title, short feature movie);
- Information related to the usage of the content (copyright pointers, usage history, broadcast schedule);
- Media information on the storage features of the content (storage format, encoding);
- Structural information on spatial, temporal, or spatio-temporal components of the content (scene cuts, segmentation in regions, region motion tracking);
- Information about low-level features in the content (colors, textures, sound timbres, melody description);
- Conceptual, semantic information of the reality captured by the content (objects and events, interactions among objects);
- Information about how to browse the content in an efficient way (summaries, views, variations, spatial and frequency sub-bands);
- Organization information about collections of objects and models that allow multimedia content to be characterized on the basis of probabilities, statistics, and examples;

- Information about the interaction of the user with the content (user preferences, usage history).

Until now research in this area has primarily focused on developing efficient, low-level, digital signal processing methods to extract values for image, video, and audio Descriptors such as color, shape, texture, motion, volume, and phonemes. Algorithms have been developed to automatically segment video into scenes and shots for faster browsing and retrieval or to automatically transcribe speech and video content. Multimedia metadata research is now focusing on how to automatically generate semantic descriptions of multimedia (machine recognition of objects and events) from combinations of low-level descriptors such as color, texture, and shape and audio descriptors to enable natural language querying and higher-level knowledge extraction.

Additional research efforts are investigating how to combine ontologies for specific domains, e.g., sports, medical, bio-informatics, and nanotechnology with MPEG-7 to describe multimedia content in terms relevant to the particular domain or to relate and integrate multimedia information from across domains or disciplines.

2.5.2 MPEG-21—Multimedia Framework. The goal of MPEG's latest initiative, MPEG-21 (ISO/IEC 18034–1) (Bormans & Hill, 2002), the Multimedia Framework, is to define the technology needed to support *Users* to exchange, access, consume, trade, and otherwise manipulate multimedia *Digital Items* in an efficient, transparent, and interoperable way. *Users* may be content creators, producers, distributors, service providers, or consumers. They include individuals, communities, organizations, corporations, consortia, governments, and other standards bodies and initiatives around the world. The fundamental unit of content is called the *Digital Item,* and it could be anything from a textual document or a simple Web page to a video collection or a music album.

At its most basic level, MPEG-21 provides a framework in which one *User* interacts with another *User* and the object of that interaction is a *Digital Item* commonly called content. Some such interactions are creating content, providing content, archiving content, rating content, enhancing and delivering content, aggregating content, delivering content, syndicating content, retail selling of content, consuming content, subscribing to content, regulating content, facilitating transactions that occur from any of the above, and regulating transactions that occur from any of the above.

The current MPEG-21 Work Plan consists of nine parts:

- Part 1: Vision, Technologies, and Strategies—a technical report that describes MPEG-21's architectural elements together with the functional requirements for their specification;

- Part 2—Digital Item Declaration—a flexible model for precisely defining the scope and components of a Digital Item;
- Part 3—Digital Item Identification—a specification for uniquely identifying Digital Items and their components;
- Part 4—Intellectual Property Management and Protection (IPMP)—to provide interoperability between IPMP tools, such as MPEG-4's IPMP hooks;
- Part 5—Rights Expression Language—a machine-readable language that can declare rights and permissions using the terms as defined in the Rights Data Dictionary (XrML);
- Part 6—Rights Data Dictionary—definitions of terms to support Part 5;
- Part 7—Digital Item Adaptation—adaptation may be based on user, terminal, network and environmental characteristics, resource adaptability, or session mobility;
- Part 8—Reference Software—used to test conformance with requirements and the standard's specifications;
- Part 9—File Format—this is expected to inherit many MPEG-4 concepts, since it will need to be able to encapsulate digital item information, still and dynamic media, metadata, and layout data in both textual and binary forms.

Future work plans for MPEG-21 include developing functional requirements and solutions to the persistent association of identification and description with Digital Items; scalable, error-resilient content representation; and the accurate recording of all events.

2.6 Rights Metadata

The Internet has been characterized as the largest threat to copyright since its inception. Copyrighted works on the Internet include news stories, software, novels, screenplays, graphics, pictures, usenet messages, and even e-mail. The reality is that almost everything on the Internet is protected by copyright law. This can pose problems for both hapless surfers as well as the copyright owners.

A number of XML-based vocabularies have been developed to define the usage and access rights associated with digital resources—XrML (XrML, 2003), developed by ContentGuard, and ODRL (ODRL, 2003), developed by IPR Systems are the two major contenders. XrML has been adopted by MPEG-21 as its Rights Expression Language, and ODRL was recently selected by the Open Mobile Alliance as its rights language for mobile content.

In addition there are a number of researchers investigating the development of well-defined, underlying, interoperable data models for rights management that is necessary for facilitating interoperability and the integration of information (indecs Framework, 2000; Delgado et al., 2002).

Project RoMEO (Rights MEtadata for Open archiving) (RoMEO, 2003) is investigating the rights issues surrounding the "self-archiving" of research in the U.K. academic community under the Open Archive Initiative's Protocol for Metadata Harvesting. Academic and self-publishing authors who make their works available through Open Archives are more concerned with issues such as plagiarism, corruption, or misuse of the text than financial returns to the author or publisher.

The "Indigenous Collections Management Project" being undertaken by Distributed Systems Technology Centre (DSTC), University of Queensland, in collaboration with the Smithsonian's National Museum of the American Indian, has also been investigating metadata for the rights management and protection of traditional knowledge belonging to indigenous communities, in accordance with customary laws regarding access (Hunter, 2002; Hunter, Koopman, & Sledge, 2003).

2.7 Automatic Metadata Extraction

Because of the high cost and subjectivity associated with human-generated metadata, a large number of research initiatives are focusing on technologies to enable the automatic classification and segmentation of digital resources—i.e., computer-generated metadata for textual documents, images, audio, and video resources.

2.7.1 Automatic Document Indexing/Classification. Automatic-categorization software (Reamy, 2002) uses a wide variety of techniques to assign documents into subject categories. Techniques include statistical Bayesian analysis of the patterns of words in the document; clustering of sets of documents based on similarities; advanced vector machines that represent every word and its frequency with a vector; neural networks; sophisticated linguistic inferences; the use of preexisting sets of categories; and seeding categories with keywords. The most common method used by autocategorization software is to scan every word in a document and analyze the frequencies of patterns of words and, based on a comparison with an existing taxonomy, assign the document to a particular category in the taxonomy. Other approaches use "clustering" or "taxonomy building" in which the software is pointed at a collection of documents (e.g., 10,000–100,000) and it searches through all the combinations of words to find clumps or clusters of documents that appear to belong together. Some systems are capable of automatically generating a summary of a document by scanning through the document and finding important sentences using rules like the first sentence of the first paragraph is often important. Another common feature of autocategorization is noun phrase extraction—the extracted list of noun phrases can be used to generate a catalog of entities covered by the collection.

Autocategorization cannot completely replace a librarian or information architect, although it can make them more productive, save them

time, and produce a better end-product. The software itself, without some human rules–based categorization, cannot currently achieve more than about 90 percent accuracy. While it is much faster than a human categorizer, it is still not as good as a human.

2.7.2 Image Indexing. Image retrieval research has moved on from the IBM QBIC (query by image content) system (QBIC, 2001), which uses colors, textures, and shapes to search for images. New research is focusing on semantics-sensitive matching (DCSE, 2003; Barnard, 2003) and automatic linguistic indexing (Wang & Li, 2003), in which the system is capable of recognizing real-world objects or concepts.

2.7.3 Speech Indexing and Retrieval. Speech recognition is increasingly being applied to the indexing and retrieval of digitized speech archives. Dragon Systems (Dragon Systems, 2003) has developed a system that creates a keyword index of spoken words from within volumes of recorded audio, eliminating the need to listen for hours to pinpoint information. Speech recognition systems can generate searchable text that is indexed to time code on the recorded media, so users can both call up text and jump right to the audio clip containing the keyword. Normally, running a speech recognizer on audio recordings doesn't produce a highly accurate transcript because speech-recognition systems have difficulty if they haven't been trained for a particular speaker or if the speech is continuous. However, the latest speech recognition systems will work even in noisy environments, are speaker-independent, work on continuous speech, and are able to separate two speakers talking at once. Dragon is also working on its own database for storing and retrieving audio indexes.

2.7.4 Natural Language and Spoken Language Querying. Dragon has also developed systems that allow users to retrieve information from databases using natural language queries. Such systems are expected to become more commonplace in the future (Oard, 2003).

2.7.5 Video Indexing and Retrieval. Commercial systems such as Virage (Virage, 2003), Convera (Convera Screening Room, 2003), and Artesia (Artesia, 2003) are capable of parsing hours of video, segmenting it, and turning it into an easily searchable and browsable database.

The latest video-indexing systems combine a number of indexing methods—embedded textual data, (SMPTE timecode, lineup files, and closed captions), scene change detection, visual clues, and continuous-speech recognition to convert spoken words into text. For example, CMU's Informedia project (Informedia, 2003) combines text, speech, image, and video recognition techniques to segment and index video archives and enable intelligent search and retrieval. The system can automatically analyze videos and extract named entities from transcripts, which can be used to produce time and location metadata. This metadata can then be used to explore archives dynamically using temporal and spatial graphical user interfaces, e.g., mapping interfaces or date sliders. For example—"give me

all video content on air crashes in South America in early 2000" (Ng et al., 2003).

Current research in this field is concentrating on the difficult problem of extracting metadata in real-time from streaming video content, rather than during a postprocessing step.

2.8 Search Engine Research and Development

2.8.1 Smarter Agent-based Search Engines. One of the major advances in search engines in the future will be in the use of "intelligent agents" and expert systems that apply artificial intelligence (AI), ontologies, and knowledge bases to enable all relevant information on a particular subject to be retrieved and integrated. Improved user interfaces will become available through the incorporation of expert systems into online catalog searching, i.e., "intelligent" sophisticated online systems that incorporate AI, knowledge bases, and ontologies. In the future librarians will use "intelligent agent kits" that will crawl over the Web retrieving relevant information and will analyze and interpret it to create a body of knowledge for a specific purpose. Periodic resampling will automatically keep it up-to-date. However, human intervention will still be needed to customize, supervise, and check the computer-generated results (Virginia Tech, 1997; Nardi & O'Day, 1998).

2.8.2 Federated Search Engines. Quite a large number of metadata research projects are focusing on the problems of federated searching across distributed, heterogeneous, networked digital libraries and the interoperability problems that need to be overcome (Gonçalves et al., 2001; Liu et al., 2002). For example, the MetaLib project, at the University of East Anglia, implements a single integrated environment and cross-searching portal for managing and searching electronic resources, whether these be abstracting and indexing databases, full-text e-journal services, CD-ROMs, library catalogs, information gateways, or local collections (Lewis, 2002).

2.8.3 Peer-to-Peer JXTA-based Search Engines. Peer-to-peer (P2P) search engines are based on the idea of decentralized metadata provided by networked peers rather than clients accessing centralized metadata repositories sitting on a server. Sam Joseph at the University of Tokyo has written an excellent overview of Internet search engines based on decentralized metadata (Joseph, 2003).

JXTA (short for Jxtapose) is a peer-to-peer interoperability framework created by Sun. It incorporates a number of protocols, but the most relevant to the idea of decentralized metadata is the Peer Discovery Protocol (PDP). PDP allows a peer to advertise its own resources and discover the resources from other peers. Every peer resource is described and published using an advertisement, which is an XML document that describes a network resource. JXTASearch operates over the lower-level JXTA protocols (JXTA, 2003).

Edutella (Edutella, 2002) is an RDF-based Metadata Infrastructure for P2P Applications based on JXTA. The first application developed by Edutella focuses a P2P network for the exchange of educational resources between German universities (including Hannover, Braunschweig, and Karlsruhe), Swedish universities (including Stockholm and Uppsala), Stanford University, and others.

2.8.4 Multimedia Search Engines. More and more search engines are becoming multimedia-capable—even allowing users to specify media types (images, video, or audio) and formats (e.g., JPEG, MP3, SMIL). Examples include the FAST Multimedia Search Engine (FAST, 2000), Alta Vista (AltaVista, 2003), Google Image Search (Google, 2003), Singingfish Multimedia Search (SingingFish, 2002), Friskit Music Streaming Media Search (Friskit, 2002), and the Fossick Online Multimedia and Digital Image Search (Fossick, 2003).

2.8.5 Cross-lingual Search Engines. In the future, universal translators will automatically translate a query in one particular language into any number of other languages and also translate the results into the original query language. There are a number of research projects and search engines focusing on cross-lingual search engines, e.g., SPIRIT-W3, a distributed cross-lingual indexing and search engine (Fluhr et al., 1997), and the TITAN Cross-Language Web search engine (TITAN, 2003).

2.9 Graphical/Multimedia Presentation of Results

2.9.1 Graphical Presentation of Search Results. More search engines are going to present search results in more innovative graphical ways other than simple lists of URLs. Interfaces like Kartoo (Kartoo, 2000) and WebBrain (WebBrain, 2001) illustrate the relationships between retrieved digital resources graphically. Kartoo uses Flash to provide a graphical representation of the results. The results are displayed in a 2–3D map representing sites that match your query as nodes on the map, and relationships between nodes are represented as labeled arcs. WebBrain presents search results in a graphical browse interface that allows users to navigate through related topics.

TouchGraph GoogleBrowser (TouchGraph, 2001) is a tool for visually browsing the Google database by exploring links between related sites. It uses Google's database to determine and display the linkages between a URL that you enter and other pages on the Web. Results are displayed as a graph, showing both inbound and outbound relationships between URLs.

"Friend of a Friend" or *foaf* (foaf, 2000) is an RDF vocabulary for describing the relationships between people, invented by Dan Brickley and Libby Miller of RDF Web. foafCORP (foafCORP, 2002) is an interesting semantic Web visualization of the interconnectedness of corporate America based on the foaf RDF vocabulary. It provides a simple graphical user

interface to trace relationships between board members of major companies in the United States.

2.9.2 Automatic Aggregation/Compilation Tools. The rapid growth in multimedia content on the Internet, the standardization of machine-processable, semantically rich (RDF-based) content descriptions, and the ability to perform semantic inferencing have together led to the development of systems that can automatically retrieve and aggregate semantically related multimedia objects and generate intelligent multimedia presentations on a particular topic, i.e., knowledge-based authoring tools (Little et al., 2002; CWI, 2000; Conlan et al., 2000; André, 2000).

Automatic information aggregation tools that can dynamically generate hypermedia and multimedia learning objects will be extremely relevant to libraries in the future. Such tools will expedite the cost-effective creation of value-added learning objects and will also ensure that any relevant content only recently made available by content providers will be automatically incorporated in the dynamically generated learning objects.

2.10 Metadata for Personalization/Customization

The individualization of information, based on users' needs, abilities, prior learning, interests, context, etc., is a major metadata-related research issue (Lynch, 2001a). The ability to push relevant, dynamically generated information to the user, based on user preferences, may be implemented

- either by explicit user input of their preferences;
- or learned by the system by tracking usage patterns and preferences and adapting the system and interfaces accordingly.

The idea is that users can get what they want without having to ask. The technologies involved in recommender systems are information filtering, collaborated filtering, user profiling, machine learning, case-based retrieval, data mining, and similarity-based retrieval. User preferences typically include information such as the user's name, age, prior learning, learning style, topics of interest, language, subscriptions, device capabilities, media choice, rights broker, payment information, etc. Manually entering this information will produce better results than system-generated preferences, but it is time consuming and expensive. More advanced systems in the future will use automatic machine-learning techniques to determine users' interests and preferences dynamically rather than depending on user input.

Some examples of "personalized current awareness news services" are Net2one (Net2one, 2003), MSNBC News Filters (MSNBC, 2003), and the eLib Newsagent project (eLib Newsagent, 2000). These services allow users to define their interests and then receive daily updated relevant reports. Filtering of Web radio and TV broadcasts will also be possible in the future, based on users' specifications of their interests and the embedding of stan-

dardized content descriptions, such as MPEG-7, within the video streams (Rogers et al., 2002).

2.11 Metadata for Broadband/Grid Applications

The delivery and integration of information is shifting to wireless mobile devices and high-performance broadband networks. To support research and development in advanced grid and networking services and applications, a number of broadband multigigabit advanced networks have been established throughout the world and made accessible to the research and higher education communities of these regions:

- Internet2—U.S. broadband research network (Internet2, 2003);
- GrangeNet—Australian broadband network (GrangeNet, 2003);
- Canarie—Canadian broadband network (Canarie, 2002);
- DANTE—European broadband research network (DANTE, 2003);
- APAN—Asia Pacific Advanced Network (APAN, 2003).

Related research projects are focusing on real-time, collaborative, distributed applications that require very high-quality video or high-speed access to large data sets for remote collaboration and visualization. Examples of applications include remote telemicroscopy, remote surgery, 3D visualization of large datasets (e.g., bio-informatics, astronomy data), collaborative editing of HDTV-quality digital video, and distributed real-time music and dance performances.

2.11.1 Grid Computing. Computational Grids enable the sharing, selection, and aggregation of a wide variety of geographically distributed computational resources (such as supercomputers, computer clusters, storage systems, data sources, instruments, people) and presents them as a single, unified resource for solving large-scale compute and data-intensive computing applications (e.g., molecular modeling for drug design, brain activity analysis, climate modeling, and high-energy physics) (Grid Computing, 2000). Wide-area distributed computing, or "grid" technologies, provide the foundation to a number of large-scale efforts utilizing the global Internet to build distributed computing and communications infrastructures. A list of current grid initiatives and projects can be found at http://www.gridforum.org/L_Involved_Mktg/init.htm (GGF, 2003).

2.11.2 The Semantic Grid. This term refers to the underlying computer infrastructure needed to support scientists who want to generate, analyze, share, and discuss their results/data over broadband Grid networks—basically it is the combination of Semantic Web technologies with Grid computing for the scientific community (Semantic Grid, 2003).

In particular, the combination of Semantic Web technologies with live information flows is highly relevant to grid computing and is an emerging research area—for example, the multiplexing (embedding) of live metadata

with multicast video streams raises the issue of Quality of Service (QoS) demands on the network.

Archival and indexing tools for collaborative video conferences held through Access Grid Nodes are going to be in demand. In typical access grid installations, there are three displays with multiple views. There is a live exchange of information. Events such as remote camera control and slide transitions could be used to segment and index the meetings for later search and browsing. Notes and annotations taken during the meeting provide additional sets of metadata that can be stored and shared. Metadata schemes to support collaborative meetings and collaboratories will be required.

Scientists collaborating on grid networks are going to require methods and tools to build large-scale ontologies, annotation services, inference engines, integration tools, and knowledge discovery services for Grid and e-Science applications (De Roure et al., 2001).

2.12 Metadata for Wireless Applications

Infrared detection and transmission can be used in libraries to beam context-sensitive data or applications to users' PDAs, depending on where they are physically located (Kaine-Krolak & Novak, 1995). Similarly, GPS information can be used to download location-relevant data to users' PDAs or laptops when they are traveling, e.g., scientists on field trips. Such context-sensitive applications require location metadata to be attached to information resources in databases connected to wireless networks.

The ROADNet (ROADNet, 2002) project on HPWREN (HPWREN, 2001), a high-performance wireless network, is a demonstration of the collection and streaming of real-time seismic, oceanographic, hydrological, ecological, geodetic, and physical data and metadata via a wireless network. Real-time numeric, audio, and video data are collected via field sensors and researchers connected to HPWREN and posted to discipline-specific servers connected over a network. This data is immediately accessible by interdisciplinary scientists in near–real time. Extraction of metadata from real-time data flow, as well as high-speed metadata fusion across multiple data sensors, are high-priority research goals within applications such as ROADNet.

2.13 Metadata Authentication

Manually generated metadata for Web resources cannot be assumed to be accurate or precise descriptions of those resources. The metadata and/or the Web page may have been deliberately constructed or edited so as to misrepresent the content of the resource and to manipulate the behavior of the retrieval systems that use the metadata. Basically, anyone can create any metadata they want about any object on the Internet with

any motivation. There is an urgent need for technologies that can vouch for or authenticate metadata so that Web indexing systems that crawl across the Internet developing Web index databases know when the associated metadata can be trusted (Lynch, 2001b).

Hence there are a number of research projects investigating methods for explicitly identifying and validating the source of metadata assertions, using technologies such as XML Signature. Search engines give higher confidence weightings to metadata signed by trusted providers, and this is reflected in the retrieved search results.

The XML Signature Working Group, a joint working group of the IETF and W3C (W3C XML Signature, 2003), has developed an XML compliant syntax for representing signatures of Web resources (or anything referenceable by a URI) and procedures for computing and verifying such signatures. Such signatures can easily be applied to metadata and used by Web servers and search engines to ensure metadata's authenticity and integrity. The XML Signature specification is based on Public Key Cryptography in which signed and protected data is transformed according to an algorithm parameterized by a pair of numbers—the so-called public and private keys. Public Key Infrastructure (PKI) systems provide management services for key registries—they bind users' identities to digital certificates and public/private key pairs that have been assigned and warranted by trusted third parties (Certificate Authorities).

Another approach is the Pretty Good Privacy (PGP) system (PGP, 2002) in which a "Web of Trust" is built up from an established list of known and trusted identity/key bindings. Trust is established in new unfamiliar identity/key bindings because they are cryptographically signed by one or more parties that are already trusted.

2.14 Annotation Systems

The motivation behind annotation systems is related to the issue of metadata trust and authentication—users can attach their own metadata, views, opinions, comments, ratings, and recommendations to particular resources or documents on the Web, which can be read and shared with others. The basic philosophy is that we are more likely to value and trust the opinions of people we respect than metadata of unknown origin.

The W3C's Annotea system (W3C Annotea, 2001) and DARPA's Web Annotation Service (DARPA, 1998) are two Web-based annotation systems that have been developed. Current research is focusing on annotation systems within real-time collaborative environments (Benz and Lijding, 1998), annotation tools for film/video and multimedia content (IBM VideoAnnEx, 2001; Ricoh MovieTool, 2002; ZGDV VIDETO, 2002; DSTC FilmEd, 2003), and tools to enable the attachment of spoken annotations to digital resources (PAXit, 2003) such as images or photographs.

2.15 Weblogging Metadata

Weblogging or Blogging (Sullivan, 2002; Reynolds et al., 2002) is a very successful paradigm for lightweight publishing, which has grown sharply in popularity over the past few years and is being used increasingly to facilitate communication and discussion within online communities. The idea of semantic blogging is to add additional semantic structure to items shared over blog channels or RSS feeds to enable semantic search, navigation, and filtering of blogs or streaming data.

Blizg (Blizg, 2003) and BlogChalking (BlogChalking, 2002) are two examples of Weblog search engines that use metadata to enable searching across Weblog archives and the detection of useful connections between and among blogs.

2.16 Metadata for Preservation

A number of initiatives have been focusing on the use of metadata to support the digital preservation of resources. Such initiatives include: Reference Model for an Open Archival Information System (OAIS, 2002), the CURL Exemplars in Digital Archives project (CEDARS, 2002), the National Library of Australia (NLA) PANDORA project (PANDORA, 2002), the Networked European Deposit Library (NEDLIB, 2001), and the Online Computer Library Center/Research Libraries Group (OCLC/RLG) Working Group on Preservation Metadata (OCLC/RLG, 2003).

These initiatives rely on the preservation of both the original bytestream/digital object, as well as detailed metadata that will enable the preserved data to be interpreted in the future. The preservation metadata provides sufficient technical information about the resources to support either migration or emulation. Metadata can facilitate the long-term access of the digital resources by providing a complete description of the technical environment needed to view the work, the applications and version numbers needed, and decompression schemes, as well as any other files that need to be linked to it. However, associating appropriate metadata with digital objects will require new workflows and metadata input tools at the points of creation, acquisition, reuse, migration, etc. This will demand initial effort to be made the first time a particular class of digital resource is received into a collection. However, assuming many of the same class of resource are received, economies of scale can be achieved by reusing the same metadata model and input tools.

The Library of Congress's Metadata Encoding and Transmission Standard (METS) (Library of Congress, 2003) schema provides a flexible mechanism for encoding descriptive, administrative, and structural metadata for a digital library object and for expressing the complex links between these various forms of metadata.

Other research initiatives are investigating extensions to METS to enable the preservation of audiovisual content or complex multimedia

objects such as multimedia artworks (Avant Garde, 2003; DSTC NewMedia, 2003). These approaches involve the association of ancillary and contextual information such as interviews with artists and the use of the Bit Stream Description Language (BSDL) (Amielh and Devillers, 2002) to convert objects preserved as bit streams into formats that can be displayed on the current platforms.

3. Conclusions

In this paper, I have attempted to provide an overview of some of the key metadata research efforts currently underway that are expected to improve our ability to search, discover, retrieve, and assimilate information on the Internet. The number and extent of the research projects and initiatives described in this paper demonstrate three things:

1. The resource requirements and intellectual and technical issues associated with metadata development, management, and exploitation are far from trivial, and we are still a long way from *MetaUtopia;*
2. Metadata means many different things to many different people, and its effectiveness depends on implementers resolving key issues, including:
   - Identifying the best metadata models, schemas, and vocabularies to satisfy their requirements;
   - Deciding on the granularity of metadata necessary for their needs—this will involve a trade-off between the costs of developing and managing metadata, the desired search capabilities, potential future uses, and preservation needs;
   - Balancing the costs and subjectivity of user-generated metadata with the anticipated error rate of automatic metadata extraction tools;
   - Ensuring the currency, authenticity, and integrity of the metadata;
   - Choosing between decentralized, distributed metadata architectures and centralized repositories for the storage and management of metadata.
3. Despite its problems, metadata is still considered a very useful and valuable component in organizing content on the Internet and in enabling us to find relevant information and services effectively.

References

Alta Vista. (2003). Retrieved July 28, 2003, from http://www.altavista.com/.

Amielh, M., & Devillers, S. (2002). Bitstream syntax description language: Application of XML-schema to multimedia content adaptation. WWW2002 Conference. Honolulu. Retrieved August 11, 2003, from http://www2002.org/CDROM/alternate/334/.

André, E. (2000). The generation of multimedia documents. In R. Dale, H. Moisl, and H. Somers (Eds.), *A handbook of natural language processing: Techniques and applications for the processing of language as text.* (pp. 305–327). Tampa: Marcel Dekker, Inc. Retrieved August 11, 2003, from http://www.dfki.de/imedia/papers/handbook.ps.

Artesia. (2003). Retrieved August 11, 2003, from http://www.artesiatech.com/.

Asia Pacific Advanced Network (APAN). (2003). Retrieved August 11, 2003, from http://apan.net/.

Avant Garde. (2003). Archiving the avant garde: Documenting and preserving variable MediaArt. Retrieved August 11, 2003, from http://www.bampfa.berkeley.edu/ciao/avant_garde.html.

Barnard, K. (2003). Computer vision meets digital libraries. Retrieved August 11, 2003, from http://elib.cs.berkeley.edu/vision.html.

Benz, H., & Lijding, M. E. (1998). Asynchronously replicated shared workspaces for a multimedia annotation service over Internet. *Lecture notes in computer science.* Retrieved August 11, 2003, from http://elib.uni-stuttgart.de/opus/volltexte/1999/533/.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001, May). The semantic Web. *Scientific American.* Retrieved August 11, 2003, from http://www.sciam.com/article.cfm?colID=1&articleID=00048144–10D2–1C70–84A9809EC588EF21.

Blizg. (2003). Retrieved August 11, 2003, from http://blizg.com/.

BlogChalking. (2002). Retrieved August 11, 2003, from http://www.blogchalking.tk/.

Bormans, J. & Hill, K. (2002). MPEG-21 overview V.5. Retrieved August 11, 2003, from http://www.chiariglione.org/mpeg/standards/mpeg-21/mpeg-21.htm.

Bourret, R. (2003a). XML and databases. Retrieved August 11, 2003, from http://www.rpbourret.com/xml/XMLAndDatabases.htm.

Bourret, R. (2003b). XML Database products. Retrieved August 11, 2003, from http://www.rphourret.com/xml/XMLDatabaseProds.htm.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the 7th International World Wide Web Conference (WWW7)* (pp 107–117). Brisbane, Australia. Retrieved August 11, 2003, from http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm.

Canarie. (2002). Retrieved August 11, 2003, from http://www.canarie.ca/.

CEDARS, CURL. (2002). Exemplars in digital archives. Retrieved August 11, 2003, from http://www.leeds.ac.uk/cedars/.

CIDOC CRM. (2003). CIDOC conceptual reference model. Retrieved August 11, 2003, from http://cidoc.ics.forth.gr/.

Conlan, O., Wade, V., Bruen, C., & Gargan, M. (2002). Multi-model, metadata driven approach to adaptive hypermedia, services for personalized e-learning. Second International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, Malaga, Spain, May 2002.

Convera Screening Room. (2003). Retrieved August 11, 2003, from http://www.convera.com/Products/products_sr.asp.

CORES. (2003). CORES—A forum on shared metadata vocabularies. Retrieved August 11, 2003, from http://www.cores-eu.net/.

CWI's Semi-automatic Hypermedia Presentation Generation (Dynamo) Project. (2000). Retrieved August 11, 2003, from http://db.cwi.nl/projecten/project.php4?prjnr=74.

DAML+OIL. (2001, December 18). Reference description. W3C Note 18 December 2001. Retrieved August 11, 2003, from http://www.w3.org/TR/daml+oil-reference.

DAML Ontology Library. (2003). Retrieved August 11, 2003, from http://www.daml.org/ontologies/.

DANTE. (2003). Retrieved August 11, 2003, from http://www.dante.net/.

DARPA Object Service Architecture Web Annotation Service (1998). Project Summary. Retrieved August 11, 2003, from http://www.objs.com/OSA/Annotations-Service.html.

Delgado, J., Gallego, I., Garcia, R., & Gil, R. (2002). An ontology for intellectual property rights: IPROnto. Poster at 1st International Semantic Web Conference (ISWC 2002). Retrieved August 11, 2003, from http://dmag.upf.es/flas_eng/publicaciones.htm.

Denny, M. (2002, November). Ontology building: A survey of editing tools. Retrieved August 11, 2003, from http://www.xml.com/pub/a/2002/11/06/ontologies.html.

Department of Computer Science and Engineering, University of Washington (DCSE). (2003). Object and concept recognition for content-based image retrieval. Retrieved August 11, 2003, from http://www.cs.washington.edu/research/imagedatabase/.

De Roure, D., Jennings, N., & Shadbolt, N. (2001). Research agenda for the semantic grid: A future e-science infrastructure. [Technical report]. UKeS-2002–02, UK e-Science Technical Report Series. National e-Science Centre, Edinburgh, UK. Retrieved August 11, 2003, from http://www.semanticgrid.org/html/semgrid.html.

Doctorow, C. (2001). Metacrap: Putting the torch to seven straw-men of the meta-utopia. Retrieved August 11, 2003, from http://www.well.com/~doctorow/metacrap.htm.

Dragon Systems. (2003). Retrieved August 11, 2003, from http://www.dragonsys.com/.
DSpace. (2002). DSpace durable digital depository. Retrieved August 11, 2003, from http://www.dspace.org.
DSTC FilmEd. (2003). The FilmEd project. Retrieved August 11, 2003, from http://metadata.net/filmed/.
DSTC New Media. (2003). The New Media art preservation project. Retrieved August 11, 2003, from http://metadata.net/newmedia/.
Dublin Core Metadata Initiative (DCMI). (2003). Retrieved August 11, 2003, from http://www.dublincore.org/.
Edutella. (2002). Retrieved August 11, 2003, from http://edutella.jxta.org/.
eLib Newsagent project. (1996). Retrieved August 11, 2003, from http://www.ukoln. ac.uk/services/elib/projects/newsagent/.
FAST. (2000). Multimedia Search Engine. Retrieved August 11, 2003, from http://www.multimedia.alltheWeb.com/
Fluhr, C., Schmit, D., Ortet, P., Elkateb, F., & Gurtner, K., (1997). SPIRIT-W3: A distributed cross-lingual indexing and search engine. Retrieved August 11, 2003, from http://www.isoc.org/isoc/whatis/conferences/inet/97/proceedings/A8/A8_1.HTM.
Foaf. (2000). The "Friend of a Friend" Project. Retrieved August 11, 2003, from http://www. foaf-project.org/.
FoafCORP. (2002). Retrieved August 11, 2003, from http://www.grorg.org/2002/10/foafcorp/.
Fossick. (2003). Online multimedia and digital image search. Retrieved August 11, 2003, from http://fossick.com/Multimedia.htm.
Friskit. (2002). Music streaming media search. Retrieved August 11, 2003, from http://www. friskit.com/.
GGF. (2003). Grid initiatives and projects. Retrieved August 11, 2003, from http://www.gridforum.org/L_Involved_Mktg/init.htm.
Gonçalves M. A., France R. K., & Fox, E. A. (2001). MARIAN: Flexible interoperability for federated digital libraries. In *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL2001).* Darmstadt, Germany. Retrieved August 11, 2003, from http://link.springer.de/link/service/series/0558/papers/2163/21630173.pdf.
Google. (2003). Image search. Retrieved August 11, 2003, from http://images.google.com/.
GrangeNet. (2003). Retrieved August 11, 2003, from http://www.grangenet.net/.
Grid Computing. (2000). Retrieved August 11, 2003, from http://www.gridcomputing.com/.
High Performance Wireless Research and Education Network (HPWREN). (2001). Retrieved August 11, 2003, from http://hpwren.ucsd.edu/news/011109.html.
Hunter, J. (2002). Rights markup extensions for the protection of indigenous knowledge. WWW2002 Conference. Honolulu, HI. Retrieved August 11, 2003, from http://archive.dstc.edu.au/IRM_project/paper.pdf.
Hunter, J., Koopman, B., & Sledge, J. (2003). Software tools for indigenous knowledge management. In *Museums on the Web.* Charlotte. Retrieved August 11, 2003, from http://archive.dstc.edu.au/IRM_project/software_paper/IKM_software.pdf.
IBM VideoAnnEx. (2001). Retrieved August 11, 2003, from http://www.research.ibm.com/VideoAnnEx/.
IMS. (2003). IMS Learning Resource Meta-data Specification. Retrieved August 11, 2003, from http://www.imsglobal.org/metadata/index.cfm.
indecs Framework Ltd. (2000). Retrieved August 11, 2003, from http://www.indecs.org/.
Informedia. (2003). Digital video understanding. Retrieved August 11, 2003, from http://www.informedia.cs.cmu.edu/.
Internet2. (2003). Retrieved August 11, 2003, from http://www.internet2.edu/.
IPTC. (2001). NewsML—Markup for the third millenium. Retrieved August 11, 2003, from http://www.iptc.org/site/NewsML/.
Joseph, S. (2003). Decentralized meta-data strategies. University of Tokyo. Retrieved August 11, 2003, from http://www.neurogrid.net/Decentralized_Meta-Data_Strategies-neat.html.
JXTA. (2003). Retrieved August 11, 2003, from http://www.jxta.org.
Kaine-Krolak, M., & Novak, M. (1995). An introduction to infrared technology: Applications in the home, classroom, workplace, and beyond. . . . Retrieved August 11, 2003, from http://trace.wisc.edu/docs/ir_intro/ir_intro.htm.
Kartoo. (2000). Retrieved August 11, 2003, from http://www.kartoo.com/.

Lagoze, C., & Hunter, J. (2001). The ABC ontology and model. *Journal of Digital Information, 2*(2). Retrieved August 11, 2003, from http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Lagoze/.

Lewis, N. (2002). Talking about a revolution? First impressions of Ex Libris's MetaLib. *Ariadne, 32*. Retrieved August 11, 2003, from http://www.ariadne.ac.uk/issue32/metalib/.

Library of Congress. (2002). Library of Congress: American Memory Historical Collections for the National Digital Library. Retrieved August 11, 2003, from http://memory.loc.gov/.

Library of Congress. (2003). METS (Metadata Encoding and Transmission Standard). Retrieved August 11, 2003, from http://www.loc.gov/standards/mets/.

Little, S., Guerts, J., & Hunter, J. (2002). The dynamic generation of intelligent multimedia presentations through semantic inferencing. ECDL 2002. Rome, Italy. Retrieved August 11, 2003, from http://archive.dstc.edu.au/maenad/ecdl2002/ecdl2002.html.

Liu X., et. al., (2002). Federated searching interface techniques for heterogeneous OAI repositories. *Journal of Digital Information, 2*(4). Retrieved August 11, 2003, from http://jodi.ecs.soton.ac.uk/Articles/v02/i04/Liu/.

Lynch, C. (2001a). Personalization and recommender systems in the larger context: New directions and research questions. Second DELOS Network of Excellence Workshop on Personalisation and Recommender Systems in Digital Libraries, Dublin, Ireland. Retrieved August 11, 2003, from http://www.ercim.org/publication/ws-proceedings/DelNoe02/CliffordLynchAbstract.pdf.

Lynch, C. (2001b). When documents deceive: Trust and provenance as new factors for information retrieval in a tangled Web. *Journal of the American Society for Information Science, 52*(1), 12–17. Retrieved August 11, 2003, from http://www.cs.ucsd.edu/~rik/others/lynch-trust-jasis00.pdf.

Magkanaraki, A., Karvounarakis, G., Anh, T. T., Christophides, V., & Plexousakis, D. (2002). Ontology storage and querying. Technical Report No. 308, ICS FORTH, Crete. Retrieved August 11, 2003, from http://139.91.183.30:9090/RDF/publications/tr308.pdf.

Martinez, J. (2002). MPEG-7 overview (Version 8). Retrieved August 11, 2003, from http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm.

MSNBC News Tools Home (2003). Retrieved August 11, 2003, from http://www.msnbc.com/toolkit.asp.

Nardi, B. A., & O'Day, V. L. (1998). Application and implications of agent technology for libraries. *The Electronic Library, 16*(5), 325–337.

Net2one Personalized News Informer. (2003). Retrieved August 11, 2003, from http://www.net2one.com/index2.asp

Networked European Deposits Library (NEDLIB). (2001). Retrieved August 11, 2003, from http://www.kb.nl/coop/nedlib/.

Ng, D., Wactlar, H., Hauptmann, A., & Christel, M. (2003). Collages as dynamic summaries of mined video content for intelligent multimedia knowledge management. AAAI Spring Symposium Series on Intelligent Multimedia Knowledge Management. Palo Alto, CA. Retrieved August 11, 2003, from http://www-2.cs.cmu.edu/~hdw/aaai03_ng.pdf.

Oard, D. (2003). Speech retrieval papers and project descriptions. Retrieved August 11, 2003, from http://raven.umd.edu/dlrg/speech/papers.html.

OASIS. (2003). Universal Description, Discovery & Integration (UDDI) of Web Services. Retrieved August 11, 2003, from http://www.uddi.org/.

OCLC/RLG. (2003). Preservation Metadata Working Group. Retrieved August 11, 2003, from http://www.oclc.org/research/pmwg/.

ODRL. (2003). Retrieved August 11, 2003, from http://www.odrl.net/.

Open Archival Information System (OAIS) Resources. (2002). Retrieved August 11, 2003, from http://www.rlg.org/longterm/oais.html.

Open Archives Initiative (OAI). (2003). Retrieved August 11, 2003, from http://www.openarchives.org/.

Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). (2003). Version 2.0, June 14, 2002. Retrieved August 11, 2003, from http://www.openarchives.org/OAI/openarchivesprotocol.html.

Open Archives Initiative (OAI) Registered Service Providers. (2002). Retrieved August 11, 2003, from http://www.openarchives.org/service/listproviders.html.

OpenGALEN. (2002). Retrieved August 11, 2003, from http://www.opengalen.org/.
OpenVideo. (2002). The OpenVideo project. Retrieved August 11, 2003, from http://www.open-video.org/.
PANDORA. (2002). National Library of Australia, PANDORA project. Retrieved August 11, 2003, from http://pandora.nla.gov.au/.
PAXit. (2003). PAXit image database software. Retrieved August 11, 2003, from http://www.paxit.com/paxit/communications.asp.
Pretty Good Privacy (PGP). (2002). Retrieved August 11, 2003, from http://www.rubin.ch/pgp/pgp.en.html.
QBIC. (2001). IBM's query by image content. Retrieved August 11, 2003, from http://wwwqbic.almaden.ibm.com/.
Reamy, T., (2002). Auto-categorization: Coming to a library or intranet near you! *EContent Magazine.* Retrieved August 11, 2003, from http://www.econtentmag.com/r5/2002/reamy11_02.html.
Reynolds, D., Cayzer, S., Dickinson, I., & Shabajee, P. (2002). Blogging and semantic blogging. SWAD-Europe Deliverable12.1.1: Semantic Web applications—analysis and selection. Retrieved August 11, 2003, from http://www.w3.org/2001/sw/Europe/reports/chosen_demos_rationale_report/hp-applications-selection.html#sec-appendix-blogging.
Ricoh MovieTool. (2002). Retrieved August 11, 2003, from http://www.ricoh.co.jp/src/multimedia/MovieTool/.
ROADNet. (2002). Real-time observatories, applications, and data management network. Retrieved August 11, 2003, from http://roadnet.ucsd.edu/.
Rogers, D., Hunter J., & Kosovic, D. (2002). The TV-trawler project. *International Journal of Imaging Systems and Technology,* Special Issue on Multimedia Content Description and Video Compression.
RoMEO. (2003). Project RoMEO (Rights MEtadata for Open archiving). Retrieved August 11, 2003, from http://www.lboro.ac.uk/departments/ls/disresearch/romeo.
SCHEMAS. (2002). SCHEMAS—Forum for metadata schema implementers. Retrieved August 11, 2003, from http://www.schemas-forum.org/registry/.
Semantic Grid. (2003). Retrieved August 11, 2003, from http://www.semanticgrid.org/.
Singingfish. (2002). Multimedia search. Retrieved August 11, 2003, from http://www.singingfish.com/.
SNOMED CT. The Systemized Nomenclature of Medicine. (2003). Retrieved August 11, 2003, from http://www.snomed.org/.
Sullivan, A. (2002). The blogging revolution. *Wired,* 10.05. Retrieved August 11, 2003, from http://www.wired.com/wired/archive/10.05/mustread.html?pg=2.
SUO. (2002). IEEE P1600.1 Standard Upper Ontology SUO Working Group. Retrieved August 11, 2003, from http://suo.ieee.org/.
Sure, Y., Erdmann, M., Angele, J., Staab, S., Studer, R., & Wenke, D. (2002). OntoEdit: Collaborative ontology engineering for the semantic Web. In *Proceedings of the First International Semantic Web Conference 2002 (ISWC 2002).* Sardinia, Italy. Retrieved August 11, 2003, from http://link.springer.de/link/service/series/0558/papers/2342/23420221.pdf.
TITAN. (2003). A Cross-Language WWW Search Engine. Retrieved August 11, 2003, from http://titan.mcnet.ne.jp/.
Topic Maps. (2000). Retrieved August 11, 2003, from http://www.topicmaps.org/.
TouchGraph GoogleBrowser. (2001). Retrieved August 11, 2003, from http://www.touchgraph.com/TGGoogleBrowser.html.
University of Illinois Library (UIL). (2002). University of Illinois Open Archives Collection. Retrieved August 11, 2003, from http://bolder.grainger.uiuc.edu/uiLibOAIProvider/2.0/oai.asp.
Virage. (2003). Retrieved August 11, 2003, from http://www.virage.com/.
Virginia Tech. (1997a). Digital libraries and software agents. Retrieved August 11, 2003, from http://scholar.lib.vt.edu/digilib/reports/agents.pdf.
Virginia Tech. (1997b). Ontologies and agents in digital libraries. Retrieved August 11, 2003, from http://ei.cs.vt.edu/~cs6604/f97/agents.htm.
W3C Annotea Web Annotation Service. (2001). Retrieved August 11, 2003, from http://annotest.w3.org/.
W3C RDF Syntax and Model Recommendation. (1999). Retrieved August 11, 2003, from http://www.w3.org/TR/REC-rdf-syntax/.

W3C RDF Vocabulary Description Language 1.0. (2003). RDF Schema, W3C working draft. Retrieved August 11, 2003, from http://www.w3.org/TR/rdf-schema/.

W3C semantic web activity. (2002). Retrieved August 11, 2003, from http://www.w3.org/2001/sw/Activity.

W3C Web Ontology Language (OWL). (2003). Guide, version 1.0, W3C working draft. Retrieved August 11, 2003, from http://www.w3.org/TR/owl-guide/.

W3C Web Ontology (WebOnt) Working Group. (2003). Retrieved August 11, 2003, from http://www.w3.org/2001/sw/WebOnt/.

W3C Web Services Activity. (2003). Retrieved August 11, 2003, from http://www.w3.org/2002/ws/.

W3C Extensible Markup Language (XML). (2003). Retrieved August 11, 2003, from http://www.w3.org/XML.

W3C XML Protocol Working Group. (2003). Simple object access protocol (SOAP). Retrieved August 11, 2003, from http://www.w3.org/2000/xp/Group/.

W3C XML Query. (2003). Retrieved August 11, 2003, from http://www.w3.org/XML/Query.

W3C XML Schema Language. (2003). Retrieved August 11, 2003, from http://www.w3.org/XML/Schema.

W3C XML Signature Working Group. (2003). Retrieved August 11, 2003, from http://www.w3.org/Signature/.

Wang, J. Z., & Li, J. (2003). Evaluation strategies for automatic linguistic indexing of pictures. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. Barcelona, Spain.

WebBrain. (2001). Retrieved August 11, 2003, from http://www.Webbrain.com/.

Web Services Description Language (WSDL). (2003). Version 1.2, W3C working draft. Retrieved August 11, 2003, from http://www.w3.org/TR/wsdl12.

XrML. (2003). Retrieved August 11, 2003, from http://www.xrml.org/.

ZGDV VIDETO. (2002). ZGDV video description tool. Retrieved August 11, 2003, from http://www.rostock.igd.fraunhofer.de/ZGDV/Abteilungen/zr2/Produkte/videto/index_html_en.

Can Document-genre Metadata Improve Information Access to Large Digital Collections?

KEVIN CROWSTON AND BARBARA H. KWASNIK

ABSTRACT

WE DISCUSS THE ISSUES OF RESOLVING the information-retrieval problem in large digital collections through the identification and use of document genres. Explicit identification of genre seems particularly important for such collections because any search usually retrieves documents with a diversity of genres that are undifferentiated by obvious clues as to their identity. Also, because most genres are characterized by both form and purpose, identifying the genre of a document provides information as to the document's purpose and its fit to the user's situation, which can be otherwise difficult to assess. We begin by outlining the possible role of genre identification in the information-retrieval process. Our assumption is that genre identification would enhance searching, first because we know that topic alone is not enough to define an information problem and, second, because search results containing genre information would be more easily understandable. Next, we discuss how information professionals have traditionally tackled the issues of representing genre in settings where topical representation is the norm. Finally, we address the issues of studying the efficacy of identifying genre in large digital collections. Because genre is often an implicit notion, studying it in a systematic way presents many problems. We outline a research protocol that would provide guidance for identifying Web document genres, for observing how genre is used in searching and evaluating search results, and finally for representing and visualizing genres.

Kevin Crowston and Barbara H. Kwasnik, Syracuse University School of Information Studies, 4–206 Centre for Science and Technology, Syracuse, NY 13244–4100
LIBRARY TRENDS, Vol. 52, No. 2, Fall 2003, pp. 345–361

INTRODUCTION

Current computerized information-access systems face a fundamental limitation: they know what documents say but not what they mean or for what purposes they might be useful. Extracting and representing the meaning of documents is difficult and time consuming, and automatic systems still have significant limitations. We note, though, that humans rarely have to read every word of a document to understand its purpose. Instead, people take a shortcut: they start by identifying the kinds of documents they are faced with (i.e., the document's genre), and then use different types of documents in appropriate ways. For example, a grant proposal is used differently from a syllabus, a product brochure, or a bank statement. Accordingly, differences in an information situation are often reflected in the *kind* of document that is considered helpful (e.g., a problem set, a lesson plan, and a tutorial about mathematics are all about math but useful in different situations). Information-access systems would be more useful for many tasks if they could similarly distinguish the purpose of documents and handle them in appropriate ways.

In this paper we discuss the possibility of improving information access in large digital collections through the identification and use of document genre as a facet of document and query representation. First, we provide some historical background on the concept of genre and the approach it provides to the problem of incorporating context into information retrieval. We outline the framework of the information-retrieval problem with respect to genre and some traditional resolutions that have been attempted. Finally, we outline a research agenda that addresses some of the questions and issues that investigating genre entails.

THEORY: DOCUMENT GENRE

Rhetoricians since Aristotle have attempted to classify communications with similar form or purpose into types or "genres." Numerous definitions of genre, or discourse type, have been suggested (e.g., Longacre, 1983; Miller, 1984; Swales, 1990). In our discussion, we draw on the definition of genre proposed by Orlikowski and Yates (1994), who describe genre as "a distinctive type of communicative action, characterized by a socially recognized communicative purpose and common aspects of form" (p. 543). For instance, this document is an example of the journal article genre. It has a form familiar to most researchers and practitioners and is monitored by the journal's editorial policies as well as the profession's communication practices. There are many document genres: some common, such as a report or a newsletter, and others restricted to specific domains, such as the course syllabus or a problem set in higher education. Genre is applicable to electronic as well as physical documents. For example, in a study of Web documents, Crowston and Williams (2000) were able to identify documents of many familiar genres and of a few genres that seemed to

be new to the Web, such as the home page (Dillon & Gushrowski, 2000) or the hotlist.

Genre is useful because it makes documents more easily recognizable and understandable to recipients, thus reducing the cognitive load of processing them (Bartlett, 1932 [1967]). Yates and Sumner (1997) argue that, on the Web genres help in both the production and consumption of documents because genre adds "fixity" in a medium that does not otherwise distinguish very well between text types (say, a book and a Post-it). In our preliminary studies of people searching the Web (Roussinov, Crowston, Nilan, Kwasnik, Liu, & Cai, 2001), we observed that the genre of the document was one of the clues used in assessing document relevance, value, quality, and usefulness.

The Problems of Information Access

To explain how genre can be useful, we will first briefly review the problems faced by an information-access system. An information-access system has three components: 1. the users, who approach the system with contextually based information needs; 2. a store of information (e.g., the documents or databases); and 3. an intermediating mechanism to connect needs and information. The intermediary may be a person, a search algorithm, a browsing environment, or a summarizer, among others.

The basic process of matching users' needs to potentially useful information in the system is complicated by many factors. First, problems may occur due to improper or incomplete representations of the information itself. When the information-access system is created, the documents or texts must be represented in such a way that they can be retrieved again as needed. Librarianship has occupied itself for over a century with systematic approaches to organizing and representing information in systems. In creating bibliographic records, we call this process cataloging; in organizing actual documents or topics for meaningful retrieval, we call it classification; in providing access to bibliographic databases, we call the representation process indexing.

There are similar processes of information representation on the Web and in many other applications in which large stores of information are prepared for eventual use in the future. Many of the schemes are adaptations of traditional schemes, such as that used on Yahoo.com, the Dublin Core Project, or the GEM Metadata Project for educational materials (http://www.thegateway.org/). Others comprise grassroots, emergent sorts of organization and representations, such as the evolving classification on eBay.com or amazon.com (Kwasnik & Liu, 2000; Kwasnik, 2002). An increasingly popular approach relies solely on the full text of the documents.

Another problem that may arise is that the process itself of matching users' queries to the document representations may be inadequate or

faulty. Much effort on the part of information scientists has been spent in developing and perfecting search strategies, including various matching algorithms, probabilistic techniques, citation mapping, and natural language processing. These efforts struggle with many obstacles—among them the difficulties of evaluating search results in real environments, as well as problems of scaling, reliability, and the representation bottleneck.

On the user side, we have people in need of information. Often, though, users are unable to precisely specify what it is they need and, even if they do, the ways in which humans articulate their needs produces a great variety of expression. The problem of appropriate representation of users' queries is not just a question of finding the correct representation according to some absolute criteria. Because information use is situated in specific contexts, there is also the need to be able to represent the information in such a way that a match can be made not only on the level of physical description and topic, for example, but also in terms of matching the information with a potential use. For example, consider a person approaching a system with the query, "I want to prepare a Passover dinner." At a certain level we can see that there is a need for concrete information in the form of actual recipes. We might even interpret this as a "known item search." However, recipes may satisfy the need only partially, since the person may want to know much more about the rituals and meanings of a Passover dinner and not just the food itself. The information need may be either broader than what is asked for or much narrower and more specific. Furthermore, we know that people ask for what they expect they can get that most closely matches what they *really* want, and thus their requests are often presented in a compromised form.

Thus, we can see that topic alone is not enough to define an information problem because different users may require different solutions to seemingly similar information problems. Indeed, even the same user may require different information at different times. These different needs arise because the situation (or context) of a user determines not only what topics are requested and what strategies are invoked in searching and evaluating output but also what types of resources are considered relevant and useful. While we know that it is important to understand the situation of the user, the representation of the situation and then its implementation in a system is a difficult problem. Our efforts to create user profiles, universal situation grammars, and so on suffer from limitations of scope to specific domains and lack of extensibility and flexibility.

Why We Think Identification of Genre Would Be Useful

We suggest that enhancing document representations by incorporating nontopical characteristics of the documents that signal their purpose—that is, their genre—will enrich document (and query) representations in such a way that they resonate more truly with the information need of a user as situated in a particular context.

Because most genres are characterized by both form and purpose, identifying the genre of a document provides information as to the document's purpose and its fit to the user's situation, which can be otherwise difficult to assess. For instance, a university professor looking for information about computer database systems for the class that she teaches would most likely be interested in documents of educational genres (e.g., syllabi, assignments, class notes). On the other hand, when working on a research paper in the database area, the same professor would more likely appreciate scholarly work (e.g., papers, annotated bibliographies, calls for papers). The relevant documents for these two searches would be quite different, even though the topic and query keywords might be nearly the same.

Explicit identification of genre seems particularly important for large digital collections because any search of these collections usually retrieves documents with a diversity of genres and, what is worse, these genres are undifferentiated by obvious clues to their identity. This is in contrast to nondigital information-seeking situations in which the searcher generally has an idea of what sorts of documents exist in the collection. Even if he or she does not, clues of physical form and location increase the chances that a document's genre is recognized. For example, a user searching in a library can visually distinguish CDs from monographs, encyclopedias, or newspapers. Similarly, a user searching a database containing only journal articles has already implicitly restricted the search to that genre of documents. On the Web, however, a search of a large and diverse document collection will usually retrieve some documents of relevant genres along with many documents of irrelevant genres—a low-precision result—even if all retrieved documents conform to search specifications regarding the topical content of the document.

Recognition of genre also has implications for automated methods of representing documents, such as automated summarization and indexing. A one-size-fits-all approach to summarizing or evaluating Web documents without regard for their form and function is likely to misrepresent many of them. For example, a newspaper article can be summarized by the first few sentences of the document, but such an approach will not work for a home page or a frequently-asked-questions document (FAQ) (Marcu, 1997). When medical information is sought, it makes a difference to the evaluation whether the document retrieved is a newsletter, a personal home page, or a hospital's patient information site.

How Librarians Have Addressed the Notion of Genre in Library Information Systems

We do not mean to imply that information science has never addressed the notion of genre, or that genre has not been incorporated into any information-representation schemes. Indeed, several classification systems allow some articulation of genre, and many metadata standards, including

the Dublin Core, include a field for genre. The treatment of genre is limited or not very well defined, however. Our understanding of the nature and role of document genre is still nascent, and so the use of this kind of information is underdeveloped in information-retrieval systems. Furthermore, it is not clear whether the extension of genre designations originally designed for physical collections will export well to digital ones.

Historically, most library information systems took genre for granted since most collections contained only a limited array of document types. The exceptions are literary genres (such as poetry) and publication types (such as almanacs or newspapers), which have had a lively existence in explicit document representation for several centuries. Aside from these, the primary facets of access to documents in traditional systems are descriptive components and subject, while genre is relatively rare. The descriptive access points derive from traditional ways of talking about books and book-like documents, and include: title, author, place and name of publisher, edition, date, series, physical description in terms of pages, size, volumes, and sometimes information about components, which are called *analytics*. The subject analysis of a document captures what a document is about—that is, its topic.

Librarians and information scientists have recognized that the topical approach is extremely important but insufficient in some situations and completely inappropriate in others. Not every document is necessarily about something. Sometimes the document's nature as a document represents the most important or useful aspect of it. For instance, on the one hand we can say that a book may be *about* symphonies—their history or structure—but what is Beethoven's Fifth *about*? It simply is. A symphony has a form and identifiable characteristics but it does not have a readily identifiable topic, *per se,* except that which can be attributed to it through subtle and nonconsensual processes of interpretation. As the notion of *document* becomes broader and more diverse, as it does in the environment of the Web, we can see how the concept of subject does not stretch very well to cover all types of information.

In response to the need to identify a document's form or genre in addition to its subject, librarians have created auxiliary tools in the form of tables and subdivisions to be used with existing topically based classification and subject-heading schedules. Here are a few examples:

The *Dewey Decimal Classification (DDC)* (Dewey, Mitchell, Beall, Matthews, & New, 1996) provides several ways to denote a document's form or genre. The first is to incorporate a designation in the number itself. This is used in the 800s, which cover *belles lettres*. The first part of the number designates country/language, and the final digits represent the genre—1 for poetry, 2 for drama, 3 for fiction, 5 for speeches . . . 7 for humor and satire, and so on (Table 1).

These genre designations are limited to the genres generally accepted by Western literary scholars and do not necessarily do a good job of

Table 1.

English poetry	821
English drama	822
English fiction	823
English speeches	825
English humor and satire	827
Bulgarian poetry	891.81
Bulgarian drama	891.82
Bulgarian fiction	891.83
Bulgarian speeches	891.85
Bulgarian humor and satire	891.87

describing emerging, culturally diverse, or hybrid genres. Still, it is a way of privileging genre in the organization of literary works. It is interesting to note, however, that most public libraries do not make use of this formal system for fiction and arrange such works by author, with the ad hoc tradition of separating out popular genres into separate sections for easy access and browsability: Mysteries, Romances, Science Fiction, and so on.

Another technique in the *DDC* is to use suffixes from the Tables. The number for the topic is established, and then suffixes from some Table are added to denote the form or genre. For example:

Table 2.

Middle Eastern Cooking	641.5956
Middle Eastern Cooking Encyclopedia	641.5956+03 (dictionaries & encyclopedias)
Middle Eastern Cooking Magazine	641.5956+05 (serial publication)

In physical collections, the suffixes serve to distinguish materials on the same topic but in different publication formats. This notion of form/genre evolved from the physical distinctions of publication and document types and thus is grounded in publishing practices and realities. The further interpretation of how such documents will be used remains implicit in the nature of the forms themselves but has practical implications for collections. For instance, many dictionaries and encyclopedias comprise the noncirculating reference collection, magazines are indexed and stored differently than are books, and so on. In terms of digital collections, however, where the physical clues of publication format are largely absent, these suffixes might provide useful indicators for sorting and filtering search results.

Another way in which subject is indicated on the bibliographic record is through the use of subject terms from a thesaurus or list, such as the

Library of Congress Subject Headings (LCSH). The LCSH comprises an evolving list of terms used by catalogers to assign subject designations to a work. Terms can denote topics, such as "sonnets," in which case this would be a work *about* sonnets, not the sonnets themselves. Proper names may also be subjects. For example, a document *about* William Shakespeare will be assigned Shakespeare's name as a subject, while a work *by* William Shakespeare would not. Modern cataloging practices abound in confusions about topic, creative responsibility, and genre/form, since in many documents these three are inextricably fused. This confusion extends to searchers as well, who do not realize that searching for a genre using *LCSH* is problematic at best.

This distinction of reserving subject headings for topics/subjects only is somewhat moderated by the addition of a subdivision. There are several kinds of subdivisions that can be used to "subdivide" a subject by time, geographical location, and further topical aspects. For example:

Witchcraft—Sweden
Witchcraft—15th Century
Witchcraft—Biblical teaching

The subdivisions of interest here, though, are the ones from the Form Subdivisions list. This type of subdivision allows the cataloger to further describe a work by its form or literary genre. This list is limited to several hundred well-established types. The genres included have literary warrant, since every subject heading and division in the *LCSH* was developed for an existing, rather than a hypothetical, work.

Witchcraft—Bibliography
Witchcraft—Case studies
Witchcraft—Dictionaries
Witchcraft—Handbooks, manuals, etc.
Witchcraft—Periodicals
Witchcraft—Poetry

The fact remains, however, that form and genre are not, as a rule, an important finding aid in traditional systems. For instance, the work *Final Environmental Impact Statement for the Green Mountain and Finger Lakes National Forests Land Resource Management Plan* is assigned the following subject headings from the *LCSH:*

Forest reserves—Vermont Green Mountain National Forest
Forest reserves—New York (State)—Finger Lakes National Forest
Forest management—Vermont Green Mountain National Forest
Forest management—New York (State)—Finger Lakes National Forest

Green Mountain National Forest (Vt.)
Finger Lakes National Forest (N.Y.)

Thus, this work can be retrieved by the names of either of the two national forests covered in the report and by two topics: *forest reserves* and *forest management.* It is not possible to retrieve this work as an *environmental impact statement* except for the coincidence that the terms appear in the title and would come up on a keyword search. There are many genres such as this one that serve a useful purpose as templates and are of interest in their own right, aside from the specific topic, but since this work is an example of an environmental impact statement, rather than a work about one, there is no subject heading assigned for this important aspect of the document.

Some libraries recognize that genre and form are often perceived as "topical" and have made some additional access points to accommodate this. For instance, the Rare Book, Manuscript, and Special Collections Library at Duke University (http://scriptorium.lib.duke.edu.genre-headings.html) has an interesting set of auxiliary tools for searching its collection. One of these is a Genre/Form list from which genre terms can be used for searching as if they were topics. Here is a sample of terms from that list:

Accounts
Business letters
Manuscripts
Official reports
Pattern books
Petitions
Recipes
Seals
Subliterary papyri
Tax returns
Vouchers

It is immediately obvious how very helpful such a list might be in studying the communicative forms of the cultures represented in the collection.

How to Study Genre

Having presented our case for understanding more about document genres in order to enhance retrieval of information from large digital collections, we turn now to the issues of precisely how we might study this phenomenon. Because genre is often an implicit and subtle notion, studying it in a systematic way presents many problems. Our overarching question is, Would identification of document genres improve information access technologies in large digital collections such as digital libraries and the

Web? This question cannot be answered directly, given the current understanding of genre or of genre's role in information retrieval. Thus, we envision a research agenda for investigating genre that proceeds through a series of componential studies, each of which we see as necessary for a full understanding. Thus, in answering the central question with respect to genre, it is necessary to investigate the following:

- The identification of Web document genres from the users' perspective and articulated in the users' own terms;
- The creation of a faceted (i.e., multidimensional) classification of these genres that can be used for controlled investigations in later stages of study;
- An investigation of how users integrate genre metadata into their own searching, evaluation, and use of documents;
- An evaluation of the degree to which incorporation of genre metadata in information-access systems makes a difference to the effectiveness of searching, sorting, ranking, and eventual use of documents; and
- An evaluation of various interfaces for visualizing and presenting genre metadata once it has been identified.

We also recognize that studying genre cannot be a once-and-for-all endeavor, since new genres are emerging all the time and old ones are being used in ways that are different than originally conceived. Thus, we propose that any study of genre must also establish a conceptual framework from within which to design continuing investigations. That is, we need a set of working hypotheses based on what we know about genre as a social construct. How are genres recognized? How do they evolve and change? How are they used and understood?

Studying Genre from the Users' Perspective

We take it as a given that studies of genres must be based in real situations with real users. Since traditional designations of document genres will probably not adequately or accurately describe all the document types present and emerging on the Web and in digital libraries, it would make no sense to use such designations as a checklist against which digital genres are compared. Such a comparison would inevitably miss genres new or unique to the Web or, even more confusing, mistake traditional genres that have been adapted to new uses on the Web.

Thus, we see the first step as a descriptive phase of inductively extracting from what people say about their terminology and sense of genres. At the same time, we recognize that studying the entire range of possible document genres and the tasks for which identification might be useful is not realistic. Furthermore, genres can be specific to a particular discourse community, so

too broad a scope may make it difficult to identify a useful set of genres in a manageable time with limited resources. Thus, as a first step we suggest that document genres be studied for a particular set of users, such as lawyers, educators, city planners, real estate agents, and so on.

If the right population is chosen, limiting the scope does not necessarily mean that only a small subset of genres will emerge. For instance, teachers can potentially search for a wide range of topics and utilize documents of many genres, including a number that are particular to education, e.g., "lesson plans" and "academic standards." This diversity means there would be a wide range of potential document types to study. On the other hand, in the domain of education there is a core set of tasks that provides a base on which to study the impact of genre identification. Such tasks and situations for teachers include, for instance, writing a lesson plan, creating a reading list, adapting an existing class, or developing tests and assignments. For administrators it might include conforming to educational standards, managing human resources, writing reports, and communicating with students and parents. Limiting the domain of inquiry will help focus a study, establish a reasonable scope, and provide a manageable set of situations with which to work and on which to test the impact of genre identification.

We realize that limiting a study in this way also limits the ability to generalize, but initially the aim is to show that genre identification is of value for certain tasks. Having demonstrated the basic concept for educators, for instance, we expect that it would be possible to then extend the principles beyond the domain of education.

Identification of Genres

In order to design effective empirical studies to investigate people's use of genre, it is necessary to identify, describe, and categorize the range of document genres used by the target population and the tasks associated with these documents. There is a substantial body of work on analyzing genre in printed documents and some work studying them on the Web (e.g., Bretan, Dewe, Hallberg, & Wolkert, 1998; Crowston & Williams, 2000; Dillon & Gushrowski, 2000; Furuta & Marshall, 1996; Karlgren, Bretan, Dewe, Hallberg, & Wolkert, 1998; Stamatatos, Fakotakis, & Kokkinakis, 2000). However, these studies have typically been top-down, that is, they analyzed a set of documents based on theoretical principles or according to a priori classifications. For example, Crowston and Williams (2000) based their classification on the *Art and Architecture Thesaurus* (Petersen, 1994), and a number of studies used the categories of the Brown Corpus.

A top-down approach to genre is problematic for two reasons. First, genres are socially constructed, so different social groups using documents with similar structural features may think about them and describe them

differently. A document may be unfamiliar and difficult to understand for someone outside of the community in which the genre is used. Therefore, it is important to capture the users' own language and understanding of these genres. Second, it is imperative to extend any investigation to genres that are not necessarily vetted by traditional schemes, such as those that come out of domain-specific work (e.g., "block-scheduled curriculum plans"). As pointed out by Dillon and Gushrowski (2000, p. 202), genres are no longer necessarily "slow-forming, often emerging only over generations of production and consumption." Thus, we assume that a traditional typology of genre or document forms will not be sufficient to describe the emerging and dynamic genres identifiable by users in general and our community in particular.

Some researchers have attempted to identify genres bottom-up through relatively small-scale user studies (e.g., Dewe, Karlgren, & Bretan, 1998; Nilan, Pomerantz, & Paling, 2001). However, we do not as yet have a fully articulated set of data that reveals what genres people recognize nor for what tasks they find documents of specific genres useful.

In investigating the range of genres identified by users, we suggest that the following questions should be addressed:

- How do people talk about the genre of documents?
- Does the naming and identification of digital-document genres correspond to the naming of traditional nondigital genres?
- How do people understand and make use of new, unnamed, emerging, and "colonized" genres (Beghtol, 2000) in digital collections?
- What clues do people use to identify genre when engaged in information-access activities?
- What facets (basic attributes) of genre do people perceive?

Creation of a Faceted (i.e., Multidimensional) Classification of Genres

If genre is to be used as another facet of description for documents and queries, we are still left with the issue of how to describe genre itself in such a way that it can be implemented in a system. Genre itself is a multidimensional phenomenon, incorporating form, function, and the numerous clues and components that allow us to discriminate one genre from another. Toward this end we see the need to create a rich and flexible description of document genres that will do justice to their complexity while at the same time providing a structured tool for systematic inquiry. One way to achieve this is through a faceted classification.

A classification will help determine the level of granularity that can be achieved in genre identification. Genre complexity can be managed by organizing the genres in a classification from more general to more specific. By picking appropriate levels of specificity it might be possible to

avoid having to identify hundreds of detailed genres, while still providing a basic level of distinction in areas of particular interest.

Most organized lists of genres are structured as hierarchies. The criticism of traditional hierarchies is that they rely on a single organizing principle, which may not be useful for all cases. To overcome this problem we suggest using the facet analysis approach. Facet analysis identifies *multiple* fundamental dimensions along which objects, such as genres, can be described and then clustered. For example, a genre such as a "lesson plan" can be identified by its source, its purpose, its structural features, and so on. Each facet, or basic dimension, can be articulated following its own logic and subsequently can be used for its own type of clues for classification. In suggesting the use of facet analysis we follow the example of previous genre-identification studies such as Päivärinta (1999), Tyrväinen and Päivärinta (1999), and Karjalainen, Päivärinta, Tyrväinen, and Rajala (2000), who looked at the management of enterprise documents, and Kessler, Nunberg, and Schuetze (1997), who sought to identify a limited set of facets for communicative purposes.

A faceted approach to classification is pragmatic and not dependent on any one conceptual perspective. It allows for the development of description and clustering using a number of fundamental dimensions, rather than just one. The results of this process yield a classification that is flexible, expressive, and hospitable to new genres and genre combinations. It also allows a view of genres at a variety of conceptual levels, from the general and inclusive to the very specific, which will be useful in simulations in later phases of inquiry.

How would a faceted approach to genres work? In principle, this approach requires several passes. The first pass identifies and labels facets that seem to be important. These might include form, content, style, implied use, and the relationship of that document to others. These facets serve as starting points, and new facets may emerge. After identifying the basic facets, one must again review the entire corpus repeatedly to see the range of categories on which these facets are revealed—for instance, what do people use to describe "source"? The process continues until saturation is reached (i.e., no new categories emerge). If necessary, more data are collected.

Once the Web genres are identified, it might be interesting to compare them to the more traditional sources of genres for overlaps of structure and coverage. We expect that there will be a significant amount of redundancy among the genres identified in this way. The aim is to generate a classification that reflects not only currently identified genres but will also flexibly accommodate identification of emerging and future genres (Beghtol, 2000; Kwasnik, 1999), thus providing a basis for future work in this area.

How Users Integrate Genre Metadata into Their Own Searching, Evaluation, and Use of Documents

Besides identifying genres and their attributes, any study of genre effectiveness must also establish how people in fact utilize genre in searching, in generating queries, and in their evaluation of documents. To accomplish this, further observations are necessary in order to answer the following questions:

- In what contexts and for what tasks is the identification of genre useful?
- To what extent are documents of various genres specific to certain tasks as opposed to being generally useful?
- To what extent are people interested in documents of genres specific to their domain and environment vs. those used more widely?

In summary, the results of the necessary initial phases of studying genres on the Web would be a better understanding of how users in the community of interest describe the genre of documents and how they use genre when they work with documents to solve information problems. This phase should also provide a database of documents categorized according to their user-specified genre, genre features, user evaluations, and related information tasks. Such a database can be used as the baseline for simulations and evaluation studies in subsequent phases. It would also provide an inductively derived faceted classification scheme for usefully clustering genres and features and for determining granularity.

Evaluating the Effectiveness of Genre Identification

Using the basic information discovered in the preliminary phases, it should then be possible to carry out controlled user studies whereby various aspects of genre use can be manipulated to see the differences in retrieval effectiveness associated with each manipulation. Thus, in this phase one can study how genre metadata can best be utilized in information-access tasks. By "best" we mean initially improving users' performance (e.g., time, accuracy, or perceived usability) in information searching, filtering, and evaluation tasks. Ultimately, of course, "best" means improving the performance in the kinds of information tasks people face in their day-to-day work lives.

This phase in the research plan must address several questions:

- How best to use genre metadata in information-access systems?
- To what extent does providing genre metadata improve performance and utility?
- Which specific facets of genre improve performance most?
- To what extent does using genre metadata to cluster and/or rank documents improve performance and utility?

- Can genre metadata be used to inform other aspects of the process of searching, summarizing, or evaluating documents?

To answer these questions, one approach would be to develop simulated information-access system interfaces that incorporate genre metadata and then to conduct evaluations of the efficacy of these interfaces in controlled laboratory experiments. Depending on resources and available design expertise, one could implement these prototypes in various ways, starting with storyboards and paper mockups, scripted interfaces and, if possible, by implementing them on a real search engine. The following are some suggestions for scenarios in which genre information could be implemented in order to test its efficacy:

- Provide aids for query construction using genre metadata;
- Cluster documents based on genre metadata (explicitly labeled vs. labeled using other techniques vs. unlabeled);
- Show genre information in combination with other metadata;
- Raise or lower a document's rank based on genre metadata; or
- Incorporate genre-specific information-access processes.

The method of presentation of genre information will inevitably influence its efficacy. For this reason, before any retrieval evaluation can take place, this aspect of genre must also be investigated. Specifically, studies should investigate:

- How best to represent and display genre metadata to the user and receive feedback?
- What level of granularity of genre metadata improves performance most (i.e., how specific should the description of genre be)? Is there, perhaps, a basic level of genre that is neither too general and abstract, nor too specific?
- To what extent does combining genre metadata with other kinds of metadata (e.g., subject) improve performance? For example, is it more useful to identify documents as "fifth-grade science lesson plans" or just as "lesson plans"?
- To what extent is the user's performance degraded by miscategorization of documents based on genre?

In answering these questions there are some general interface design issues that must also be addressed. For example, it is important to decide on:

- The choice between *opaque* and *transparent* modes of presentation (Roussinov et al., 2001). In transparent mode, the system will expose its identification of genre to the user by labeling or otherwise identifying

it. In opaque mode, results will make use of genre metadata, e.g., for ranking, but the user will not be made explicitly aware of it;
- If the transparent mode is used, then one must make a choice between interfaces that explicitly name genres and those that use other methods of labeling (e.g., by providing example documents).

A typical experiment might for instance contrast two interfaces. Participants could use one interface in the first half and a different interface in the second half of the experiment, with the order of presentation counterbalanced across subjects. An example task might be to find a relevant Web page for a given problem scenario.

Answering the Big Question: Does Genre Help?

Identifying Web document genres, observing how genre is used in searching and evaluation of search results, the multifaceted representation of genres, and finally the design of presentation and visualization techniques allows a systematic exploration of the overall effectiveness and utility of genre information for Web documents.

We know that people use genre information and that for many applications it is perhaps the single most important piece of information that can be provided. Nevertheless, the extent to which the general inclusion of genre information will enhance information access on the Web is an open question. There is much to understand before such information can be practically implemented. On the other hand if, as we suspect, genre information is helpful, then studying it in a systematic way, as we propose above, can provide the initial baseline understanding that must precede any automatic implementation.

References

Bartlett, F. (1932 [1967]). *Remembering: A study in experimental and social psychology.* Cambridge: Cambridge University Press.

Beghtol, C. (2000). The concept of genre and its characteristics. *Bulletin of the American Society for Information Science & Technology, 26*(5).

Bretan, I., Dewe, J., Hallberg, A., & Wolkert, N. (1998). Web-specific genre visualization. Paper presented at the WebNet 1998 Conference. Orlando, FL.

Crowston, K., & Williams, M. (2000). Reproduced and emergent genres of communication on the World Wide Web. *Information Society, 16*(3), 201–216.

Dewe, J., Karlgren, J., & Bretan, I. (1998). Assembling a balanced corpus from the Internet. Paper presented at the 11th Nordic Computational Linguistics Conference. Copenhagen, Denmark.

Dewey, M. A., Mitchell, J. S. E., Beall, J. E., Matthews, W. E. J. E., & New, G. R. E. (Eds.). (1996). *Dewey Decimal Classification and Relative Index Set,* 21st ed. Dublin: OCLC Forest Press.

Dillon, A., & Gushrowski, B. (2000). Genres and the Web: Is the personal home page the first uniquely digital genre? *Journal of the American Society for Information Science, 5*(12), 202–205.

Furuta, R., & Marshall, C. C. (1996). *Genre as reflection of technology in the World Wide Web.* [Technical Report]. College Station, TX: Hypermedia Research Lab, Texas A&M.

Karjalainen, A., Päivärinta, T., Tyrväinen, P., & Rajala, J. (2000). Genre-based metadata for enterprise document management. In *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences.* Los Alamos, CA: IEEE Computer Society Press.

Karlgren, J., Bretan, I., Dewe, J., Hallberg, A., & Wolkert, N. (1998). Iterative information retrieval using fast clustering and usage-specific genres. Paper presented at the Eighth DELOS Workshop: User Interface in Digital Libraries. Stockholm, Sweden.

Kessler, B., Nunberg, G., & Schuetze, H. (1997). Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Meeting of the European Chapter of the Association for Computational Linguistics* (pp. 32–38). San Francisco: Morgan Kaufmann Publishers.

Kwasnik, B. H. (1999). The role of classification in knowledge representation and discovery. *Library Trends, 48*(1), 22–47.

Kwasnik, B. H. (2002). Commercial Websites and the use of classification schemes: The case of amazon.com. In M. J. Lopez-Huertas (Ed.), *Challenges in knowledge representation and organization for the 21st century: Integration of knowledge across boundaries. Proceedings of the Seventh International ISKO Conference* (pp. 279–285). Würzburg: Ergon Verlag.

Kwasnik, B. H., Crowston, K., Nilan, M., & Roussinov, D. (2000). Identifying document genre to improve Web search effectiveness. *Bulletin of the American Society for Information Science and Technology, 27*(2), 23–26.

Kwasnik, B. H., & Liu, X. (2000). Classification structures in the changing environment of active commercial Websites: The case of eBay.com. In C. Beghtol, L. C. Howarth, & N. J. Williamson (Eds.), *Dynamism and stability in knowledge organization. Proceedings of the Sixth International ISKO Conference, 7*(10–13), 372–377. Toronto: Advances in Knowledge Organization.

Longacre, R. (1983). *The grammar of discourse.* New York: Plenum Press.

Marcu, D. (1997). From discourse structures to text summaries. Paper presented at the 14th National Conference on Artificial Intelligence (AAAI-97).

Miller, C. R. (1984). Genre as social action. *Quarterly Journal of Speech, 70,* 151–167.

Nilan, M. S., Pomerantz, J., & Paling, S. (2001). Genres from the bottom up: What has the Web brought us? In T. B. Hahn (Ed.), *Information in a networked world: Harnessing the flow. Proceedings of the ASIST 2001 Annual Meeting.* Washington, DC.

Orlikowski, W. J., & Yates, J. (1994). Genre repertoire: The structuring of communicative practices in organizations. *Administrative Sciences Quarterly, 33,* 541–574.

Päivärinta, T. (1999). A genre approach to applying critical social theory to information systems development. In C. H. J. Gilson, I. Grugulis, & H. Willmott (Eds.), *Proceedings of the 1st Critical Management Studies Conference, information technology and critical theory stream.* Hamilton, New Zealand: University of Waikato Management School.

Petersen, T. (1994). *Art and architecture thesaurus.* New York: Oxford University Press.

Roussinov, D., Crowston, K., Nilan, M., Kwasnik, B., Liu, X., & Cai, J. (2001). Genre-based navigation on the Web. Paper presented at the 34th Hawaii International Conference on Systems Science (HICSS-34). Maui, HI..

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. *Computational Linguistics, 26*(4), 471–498.

Swales, J. M. (1990). *Genre analysis: English in academic and research settings.* New York: Cambridge University Press.

Tyrväinen, P., & Päivärinta, T. (1999). On rethinking organizational document genres for electronic document management. In *Proceedings of the 32nd Annual Hawaii International Conference on System Sciences.* Los Alamos, CA: IEEE Computer Society Press.

Yates, J., & Orlikowski, W. J. (1992). Genres of organizational communication: A structurational approach to studying communications and media. *Academy of Management Review, 17*(2), 299–326.

Yates, S. J., & Sumner, T. (1997). Digital genres and the new burden of fixity. In *Hawaiian International Conference on System Sciences (HICCS 30).* Los Alamos, CA: IEEE Computer Society Press.

Web-based Organizational Tools and Techniques in Support of Learning

Don E. Descy

Abstract

The Internet, particularly the Web, is a wonderful free source of information that can vastly improve the array of resources available to library patrons. Unfortunately, not all information is honest and accurate, and some of it is not suitable for certain age groups. Students using the Web for research often come upon unsuitable sites. We can get around this by constructing Web learning environments containing safe sites for students. These learning experiences include WebQuests, Pathfinders, Treasure Hunts, Scavenger Hunts, and Tracks.

As librarians, we pride ourselves on providing our clients with open and free access to uncensored information. We have always been at the forefront in the fight against censorship and threats to individual privacy rights. We realize that these two civil rights are important components at the base of a free society. Granted, we do pick and choose the materials that we place in our libraries. With limited budgets and space we cannot include everything in our holdings. We do pick and choose and, in doing so, may limit materials that may be deemed "controversial" by some. Saying this, we do not limit materials and opinions just on the basis of controversy. We also try to take advantage of services that expand our holdings in the most cost-effective way we can. We form partnerships and cooperatives to purchase materials and services. We share our materials through interlibrary loan and other cooperative efforts. As good stewards of our monies and the public trust, we take advantage of as many opportunities as we can to stretch our budget and acquire accurate and unbiased information for our patrons.

Don E. Descy, Educational Technology Program Director, 313 Armstrong Hall, Minnesota State University, Mankato, MN 56001
LIBRARY TRENDS, Vol. 52, No. 2, Fall 2003, pp. 362–366

One such opportunity for expanding our horizons is the Internet in general and the World Wide Web (WWW) in particular. The WWW allows us to access amounts of information that we never dreamed possible even just a few short years ago. I remember saying years ago that the effect of the Internet on the world would be greater than that of the printing press. People thought that I was crazy back then, but it seems that this has come to pass.

The Internet is not without its problems though. Everything seems to be out there and available to our patrons with just a few keystrokes or clicks. Everything: honest information, dishonest information, biased information, and information that may be just plain inappropriate for the customer because of age or ability to understand. This can be a particular problem for librarians working in our schools. How can we have free and open access to information and yet screen this information to insure its availability in an age-appropriate form? The answer very simply is that we can't. Many schools rely on student trust that self-limits access to inappropriate sites. Many politicians and others have touted Internet filters as a panacea. Most of these individuals have good intentions.

Unfortunately, filters don't work as well as people claim. A 2001 study reported in *Consumer Reports* (2001) found that most filtering software packages failed to block one out of five undesirable sites. Many times sites that would be appropriate sources of information are denied to students. Klauck (1999) studied common Web filtering software and, using search terms common in school settings, found results that were undesirable for students.

So, what can be done? How can we give our students the opportunity to use the Web or Web-based information and yet stay in a safe environment? This is not as difficult as it may seem. We do this by guiding each step of their Web experience to assure that their keystrokes and clicks get them to the information they need with no chance of them going astray. We do this by constructing WebQuests and Pathfinders and using lists of specific sites and information that we construct or are available on the Web.

WebQuests

Bernie Dodge and Tom March first developed WebQuests at San Diego State University in 1995. A WebQuest is a self-contained, inquiry-oriented activity constructed in the form of a Web page. Some or all if the information the students will interact with comes from the Web. Information found in other library resources may also be used along with films, television, and other technologies. From this central page the student is prompted to read or view other information, and visit other Web pages constructed by the instructor, and other Web pages of supplemental information around the world. The information they gather from other sources and these Web pages is used by the students to complete their tasks.

The main page of the WebQuest contains several parts: introduction, task, process, evaluation, conclusion, and resources. Links to a separate teacher page may also be included. The *introduction*, though short in itself, is one of the most important parts of the WebQuest. This portion introduces the student to the activity. It is designed to "hook" the student and make the student want to go on. The *task* concisely states the outcome of the quest. It is usually short and to the point. WebQuests may be of short or long duration, to be completed individually or with a group. The *process* lists the steps the students will follow to achieve the desired outcome. A well-designed WebQuest is different and interesting. Students don't write reports and parrot information. Rather, they may write a play, give a presentation, have a debate or construct a final project. The fourth area is *evaluation*. Here, students see how their performance will be evaluated. Many times assessment rubrics are used. Students learn about individual and/or group grading The *conclusion* summarizes and brings closure to the activity. It also encourages students to extend their studies in or near the area covered by the WebQuest. Additional examples of projects or topics to explore may be introduced. And finally, the *resources* section may contain links to other useful sites containing supplemental and enrichment information. WebQuest etiquette allows teachers three choices. It is perfectly acceptable to use someone else's WebQuest that may be found posted on a site. Of course, permission should be asked as common courtesy. Posted WebQuests may be changed to more closely meet the needs of the curriculum. Again, it would be courteous to send a copy of this to the originator of the WebQuest. And, of course, the instructor can also make a WebQuest from scratch. Many tutorials and templates can be found with a simple Web search.

PATHFINDERS

Pathfinders are guides to information resources on a specific topic. They are designed to provide a path for students to follow that focuses on their areas of research and specifically targets the most appropriate resources available. General Pathfinders may contain print resources such as books and periodicals; nonprint resources such as videos, CD-ROMS, and audiotapes; and Internet sources such as Web sites and discussion groups. Many librarians have constructed Pathfinders or subject guides for their patrons. This is a good way to have a list of prescreened sources for students to use. As long as librarians can use a word processor, they can turn a Pathfinder into a Web page by just saving it in that form or utilizing such Web construction tools as Netscape Composer, available free as a part of the Netscape Web browser. There is even a Pathfinder for constructing Web-based pathfinders easily located on the Web (http://home.wsd.wednet.edu/pathfinders/path.htm)! Web-based Pathfinders allow our students the easy access to online resources such as Web sites, online community resources, library catalogs, encyclopedias, newspapers, and magazines.

Many of the better resources such as online databases are not easily found through the use of regular Web searching. This vast pool of virtually searchable information, called the Invisible Web (discussed elsewhere in this issue) contains some of our best resources. A simple search using the term Invisible Web will yield many lists of these resources to add to and increase the value of our Pathfinders.

Bookmarks are sometimes called "pathfinders" without a sense of direction. They are collections of seemingly unrelated sites. Many teachers use their collection of bookmarks as jumping-off points in history and English classes. A well-constructed group of bookmarks can be an extraordinary resource for teachers and students for safe surfing through troubled Web waters.

Web Treasure Hunts/ Scavenger Hunts

Web Treasure Hunts, sometimes called Scavenger Hunts, are just like regular treasure hunts except that the students use the Internet to find answers to questions. Many great sites are available to help the teacher find, construct, and utilize Treasure Hunts in their classroom (http://www.ctnba.org/ctn/k8/treasure.html, for example). Treasure hunts focus on a particular theme that a teacher in interested in using in class. They do require much more time online than WebQuests, since the students must take more time finding the sites and digesting the information. Unlike Pathfinders, Treasure Hunts may be designed to introduce students to searching and search engines to find the information they need. This may introduce variables that can't be controlled but, by using search engines designed for student use, these variables can be minimized. Other Treasure Hunts rely on carefully collected and evaluated sites and questions related to the topic under study so students will have a more controlled environment to work in. One of the fun aspects of Treasure Hunts is that they are often timed. This is a great way to keep students occupied and on task.

Tracks

A Track is a collection of sites about a similar topic. They can be used by teachers and students to create their own Scavenger Hunts or Treasure Hunts. Many times it is worthwhile for students to actually construct their own hunt. A search of many of the popular teacher Web sites will yield many lists of sites on similar subjects. These lists are a good place to get started while the teacher and students construct their own tracks and hunts.

The Web is a wonderful place. The more we use it, the more we marvel at the vast collection of information that is out there available with just the click of a mouse or the pushing of a few keys. It is also a minefield. Hazards abound. Sites with wrong, misleading, and biased information abound. Through the careful use of some of the resources that we have

discussed, it is possible to help our students navigate this minefield and find good, useful information quickly and easily.

REFERENCES

Consumers Reports. (2001, March). Digital chaperones for kids. Which Internet filters protect the best? Which get in the way? Retrieved November 21, 2002, from http://www.comsumerreports.org/Special/ConsumerInterest/Reports/0103fil0.html.

Klauk, R. R. (1999, December). Does the use of Internet filtering software in elementary schools work as intended? [Unpublished master's alternate plan paper]. Minnesota State University, Mankato.

About the Contributors

JOHN CARLO BERTOT is Associate Professor and Associate Director of the Information Use Management and Policy Institute in the School of Information Studies at Florida State University. He teaches courses in library technology planning, technology applications, and information and telecommunications policy. Bertot publishes in the areas of library management, planning, and evaluation, with a particular emphasis on library Internet use and involvement. Most recently, he coauthored *Statistics and Performance Measures for Public Library Networked Services,* through the American Library Association (2001), and coedited *Evaluating Networked Information Services,* an Information Today publication (2002). Bertot serves as editor of *Government Information Quarterly* and coeditor of *Library Quarterly.* At present, Bertot is principal investigator for a grant to develop outcomes-based assessment training modules for public libraries in Florida and co–principal investigator for a national U.S. study sponsored by the Institute of Museum and Libraries to develop training modules to assist libraries in collecting, using, and reporting network services data. Bertot serves as a U.S. delegate on the International Standards Organization's (ISO) Library Statistics committee and chair of the ISO Library Performance Indicator committee. He is also a member of the National Information Standards Organization's (NISO) planning committee for the revision of the Z39.7 *Library Statistics* standard. Additional information on Bertot and selected publications is available at http://slis-two.lis.fsu.edu/~jcbertot/.

REBECCA P. BUTLER is an Associate Professor of Instructional Technology and School Library Media in the Department of Educational Technology, Research, and Assessment at Northern Illinois University, DeKalb. She is a former school library media specialist and has also worked in public, academic, special, and medical libraries. Her areas of research interest include intellectual properties, intellectual freedom, and the history of

instructional technology. She writes a column on copyright law for *Knowledge Quest,* the journal of the American Association of School Librarians, and speaks nationally on intellectual property issues. She is currently finishing a book on copyright law.

JERRY D. CAMPBELL is the Chief Information Officer and Dean of the University Libraries at the University of Southern California, where he arrived in January 1996. In the course of his career, he has played a leadership role in numerous organizations and agencies. This has included serving as president of the Association of Research Libraries and the Triangle Universities Center for Advanced Studies, Inc. It has also included serving on the governing boards of numerous agencies such the Research Libraries Group, the Council on Library and Information Resources, the National Institute of Statistical Sciences, and the National Humanities Center. Dr. Campbell has contributed articles to books, published numerous articles in journals, and spoken and consulted widely.

KEVIN CROWSTON is an Associate Professor in the School of Information Studies at Syracuse University and Director of the Ph.D. program. He joined the school in 1996. He received his A.B. (1984) in Applied Mathematics (Computer Science) from Harvard University and a Ph.D. (1991) in Information Technologies from the Sloan School of Management, Massachusetts Institute of Technology (MIT). Before moving to Syracuse he was a founding member of the Collaboratory for Research on Electronic Work at the University of Michigan and of the Center for Coordination Science at MIT.

DON E. DESCY is a Professor in the Library Media Education Programs at Minnesota State University, Mankato and Program Director for Educational Technology. He is editor-in-chief of *TechTrends,* a journal of the Association for Educational Communications and Technology. He has written hundreds of articles, columns, and book chapters on media utilization in P–12 schools and universities. Dr. Descy has presented all over the United States and in Canada, France, Japan and China. His text on computer utilization for preservice educators is going into its fourth edition for Prentice Hall.

PATRICIA DIAMOND FLETCHER is Associate Professor in Policy Sciences at University of Maryland, Baltimore County. She teaches and conducts research on government information policy and government information system management, currently policy and implementation of electronic government. She received her doctorate and M.L.S. from the School of Information Studies at Syracuse University. Fletcher has participated in numerous national studies of information resources management

in U.S. government. In March 2001, Fletcher was conference cochair for an NSF funded workshop "Electronic Government in U.S. Cities," sponsored by the University of Illinois, Chicago. She recently completed a one-year academic fellowship at the U.S. General Accounting Office. Fletcher sits on the editorial boards of *Government Information Quarterly* and *Journal of Global Information Management.*

ADRIENNE FRANCO is the Reference and Instructional Services Librarian at the Iona College Libraries in New Rochelle, New York, where she teaches bibliographic instruction courses and coordinates library instruction and reference activities. She has created the Web site "Finding Quality Information on the World Wide Web," which was originally produced for a presentation given by her and Richard Palladino at the tenth annual meeting of the International Information Management Association in 1999. She serves as a member (and past chair) of the WALDO Information Services Committee.

JANE L. HUNTER is a Senior Research Fellow at the Distributed Systems Technology Centre at the University of Queensland. Her research interests are multimedia metadata modeling and interoperability between metadata standards across domains and media types. She was chair of the MPEG-7 Semantic Interoperability Adhoc Group, editor of the MPEG-7 Description Definition Language ISO/IEC 15838–2, and is the liaison between MPEG, W3C, and the Dublin Core Metadata Initiative.

BARBARA H. KWASNIK is a Professor in the School of Information Studies at Syracuse University in Syracuse, New York, where she has been teaching in the areas of organization of information, theory of classification, and information science. Her current research interests are in developing methodologies for studying information-related behavior and in the cognitive processes of browsing and the structure of classificatory systems, particularly the nature of unspecified term relationships. She is founder and co-coordinator of annual workshops sponsored by SIG/CR (ASIST) and coeditor or program committee member for the proceedings. Dr. Kwasnik is also active in the International Society for Knowledge Organization and is a member of the American Society of Indexers.

GREG R. NOTESS is a Reference Librarian and Professor at Montana State University-Bozeman. He has been writing, speaking, and consulting about Internet information resources since 1991. A three-time Information Authorship award winner, he is the "On the Net" and "Internet Search Engine Update" columnist for ONLINE. Notess is the author of the first three editions of *Government Information on the Internet and Internet Access Providers: An International Resource Directory.* An internationally known conference speaker

on search engines and other Internet topics, Notess has spoken at many national conferences and international meetings in Stockholm, London, Oslo, Montreal, Copenhagen, and Sydney, Australia. On the Web, Notess maintains SearchEngineShowdown.com, which reviews, compares, analyzes, and tracks news in the search engine industry.

GARY PRICE is a librarian, information research consultant, and writer based in suburban Washington, D.C. Gary is the editor and compiler of the *Resource Shelf* (http://www.resourceshelf.com), a daily electronic newsletter. A native of the Chicago area, he earned his M.L.I.S. degree from Wayne State University in Detroit, Michigan. He also holds a B.A. degree from the University of Kansas in Lawrence, Kansas. Price is a Web Search University faculty member and an inductee of the Internet Librarian Hall of Fame.

CHRIS SHERMAN is President of Searchwise, a Boulder, Colorado–based Web consulting firm, and Editor of SearchDay, a daily newsletter from SearchEngineWatch.com. He is a regular contributor to *Information Today, Online, EContent,* and other information industry journals and a regular presenter at information industry conferences and workshops. Sherman holds a master's degree in Interactive Educational Technology from Stanford University and a bachelor's degree in Visual Arts and Communications from the University of California, San Diego. He is a Web Search University faculty member and an inductee of the Internet Librarian Hall of Fame.

AMANDA SPINK is Associate Professor at the School of Information Sciences at the University of Pittsburgh. She has a B.A. (Australian National University); Graduate Diploma of Librarianship (University of New South Wales); M.B.A. (Fordham University), and a Ph.D. in Information Science (Rutgers University). Dr. Spink's research focuses on theoretical and applied studies of human information behavior and interactive information retrieval (IR), including Web and digital libraries studies. The National Science Foundation, Andrew R. Mellon Foundation, NEC, IBM, Excite, FAST, and Lockheed Martin have sponsored her research. She has published over 180 journal articles and conference papers, with many in the *Journal of the American Society for Information Science and Technology, Information Processing and Management, Interacting with Computers, IEEE Computer, Internet Research,* the *ASIST* and *ISIC Conferences.*

ANDREW G. TOROK is Professor Emeritus in the Department of Educational Technology, Research, and Assessment at Northern Illinois University, DeKalb. Formerly he taught for several years in the department of Library Science, also at Northern Illinois. He teaches classes in computer networking, online education, instructional technology, and several semi-

nars that support a large doctoral program. Dr. Torok has been active in the information industry for four decades, working as a teacher, researcher, indexer, and abstractor. He has published and presented papers nationally and internationally and served as a consultant. His research interests have included ergonomics issues relating to technology, online user studies, and communication studies. His current research interests include technology ROI and electronic learning. He also continues to engage in technical writing.

STATEMENT OF OWNERSHIP, MANAGEMENT, AND CIRCULATION

(1) Publication Title: *Library Trends.* (2) Publication Number: 0024-2594. (3) Filing Date: October 2003. (4) Issue Frequency: Quarterly (Summer, Fall, Winter, Spring). (5) Number of Issues Published Annually: 4. (6) Annual Subscription Price: $94. (7) Office of Publication Address: University of Illinois Press, 1325 S. Oak Street, Champaign, Illinois 61820; Contact Person: Cheryl Jestis, Telephone (217) 244-0626. (8) General Business Office Address: University of Illinois Press, 1325 S. Oak Street, Champaign, Illinois 61820. (9) Name and Address of Publisher, Editor, and Managing Editor: Publisher—University of Illinois Press, 1325 S. Oak Street, Champaign, Illinois 61820; Editor—Wilfred Lancaster, Library & Information Science Bldg., 501 East Daniel Street, Champaign, Illinois 61820-6211; Managing Editor—Wilfred Lancaster, Library & Information Science Bldg., 501 East Daniel Street, Champaign, Illinois 61820-6211. (10) Owner: University of Illinois Press, 1325 S. Oak Street, Champaign, Illinois 61820. (11) Bondholders: None. (12) Tax Status: Has Not Changed During Preceding 12 Months. (13) Publication Title: *Library Trends.* (14) Issue Date for Circulation Data: 6/17/2003.

(15) Extent and Nature of Circulation	Average No. Copies Each Issue During Preceding 12 Months	No. Copies of Single Issue Published Nearest to Filing Date
a. Total Number of Copies (Net Press Run)	2378.25	2605
b. Paid and/or Requested Circulation		
(1) Paid Requested Outside-County Mail Subscriptions Stated on Form 3541	1802.25	1869
(2) Paid in-County Subscriptions	0	0
(3) Sales through Dealers and Carriers, Street Vendors, Counter Sales, and Other Non-USPS Paid Distribution	0	0
(4) Other Classes Mailed Through the USPS.	0	0
c. Total Paid and/or Requested Circulation	1802.25	1869
d. Free Distribution by Mail		
(1) Outside-County as Stated on Form 3541	17.25	8
(2) In-County as Stated on Form 3541	0	0
(3) Other Classes Mailed Through the USPS	0	0
e. Free Distribution Outside the Mail	0	0
f. Total Free Distribution	17.25	8
g. Total Distribution	1819.5	1877
h. Copies not Distributed	558.75	728
i. Total	2378.25	2605
j. Percent Paid and/or Requested Circulation	99.05	99.57

(16) Publication of Statement of Ownership will be printed in the Volume 52, Number 2, Fall 2003 issue of this publication.

LIBRARY TRENDS

"*Library Trends* has become the premier thematic quarterly journal in the field of American Librarianship."
Library Science Annual

Both practicing librarians and educators use *Library Trends* as an essential tool in professional development and continuing education. They know *Library Trends* is the place to discover practical applications, thorough analyses, and literature reviews for a wide range of trends. See for yourself the breadth of topics covered in forthcoming issues.

- **THE PHILOSOPHY OF INFORMATION**
  (Winter 2004) Edited by Ken Herold
- **PIONEERS IN LIBRARY AND INFORMATION SCIENCE**
  (Spring 2004) Edited by Boyd Rayward
- **ORGANIZATIONAL DEVELOPMENT IN LIBRARIES**
  (Summer 2004) Edited by Keith Russell and Denise Stephens
- **CONSUMER HEALTH INFORMATION SERVICES**
  (Fall 2004 and Winter 2005) Edited by Tammy Mays

Institutional subscription price $100 (plus $7 for international subscribers). Individual subscription price $70 (plus $7 for international subscribers). Student subscription price is $30 (plus $7 for international subscribers). Single copies are available for $28, including postage. Order from the University of Illinois Press, Journals Department, 1325 S. Oak St., Champaign, IL 61820-6903, Telephone 866-244-0626, Mastercard, Visa, American Express, and Discover accepted.

0024-2594(200323)52:2;1-R